University of Exeter
Department of Mathematics

Statistical methods for quantifying uncertainty in climate projections from ensembles of climate models

Philip George Sansom

May 2014

Supervised by
David B. Stephenson
Christopher A. T. Ferro
Ruth E. McDonald

Submitted by Philip George Sansom, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Mathematics , May 2014.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) .....

Abstract

Appropriate and defensible statistical frameworks are required in order to make credible inferences about future climate based on projections derived from multiple climate models.

It is shown that a two-way analysis of variance framework can be used to estimate the response of the actual climate, if all the climate models in an ensemble simulate the same response. The maximum likelihood estimate of the expected response provides a set of weights for combining projections from multiple climate models. Statistical F tests are used to show that the differences between the climate response of the North Atlantic storm track simulated by a large ensemble of climate models cannot be distinguished from internal variability.

When climate models simulate different responses, the differences between the responses represent an additional source of uncertainty. Projections simulated by climate models that share common components cannot be considered independent. Ensemble thinning is advocated in order to obtain a subset of climate models whose outputs are judged to be exchangeable and can be modelled as a random sample. It is shown that the agreement between models on the climate response in the North Atlantic storm track is overestimated due to model dependence.

Correlations between the climate responses and historical climates simulated by climate models can be used to constrain projections of future climate. It is shown that the estimate of any such emergent relationship will be biased, if internal variability is large compared to the model uncertainty about the historical climate. A Bayesian hierarchical framework is proposed that is able to separate model uncertainty from internal variability, and to estimate emergent constraints without bias. Conditional cross-validation is used to show that an apparent emergent relationship in the North Atlantic storm track is not robust.

The uncertain relationship between an ensemble of climate models and the actual climate can be represented by a random discrepancy. It is shown that identical inferences are obtained whether the climate models are treated as predictors for the actual climate or vice versa, provided that the discrepancy is assumed to be symmetric. Emergent relationships are reinterpreted as constraints on the discrepancy between the expected response of the ensemble and the actual climate response,

conditional on observations of the recent climate. A simple method is proposed for estimating observation uncertainty from reanalysis data. It is estimated that natural variability accounts for 30-45% of the spread in projections of the climate response in the North Atlantic storm track.

For my wife Natalie, for all her patience, love and support.

Thank you to my supervisors, David Stephenson and Chris Ferro, for encouraging me to find my own direction, and for helping me to stay on course.

The analysis of cyclone track density would not have been possible without the assistance of Giuseppe Zappa and Kevin Hodges, who prepared the data.

My thanks to Thomas Bracegirdle for helpful conversations and example data on emergent constraints and Arctic sea ice.

Thanks also to Theo Economou, Danny Williamson, and Peter Challenor for helpful discussions.

Finally, thank you to my colleagues David Walker, Jacqueline Christmas and Zena Wood, for all the tea and biscuits that kept me going.

Contents

List of Tables

List of Figures

Publications

Material from Chapter 3 has been published in

Philip G. Sansom, David B. Stephenson, Christopher A. T. Ferro, Giuseppe Zappa, and Len Shaffrey, 2013: Simple uncertainty frameworks for selecting weighting schemes and interpreting multi-model ensemble climate change experiments. *Journal of Climate*, **26**, 4017–4037, doi:10.1175/JCLI-D-12-00462.1.

Zappa, G., L. C. Shaffrey, K. I. Hodges, P. G. Sansom, and D. B. Stephenson, 2013: A multi-model assessment of future projections of North Atlantic and European extratropical cyclones in the CMIP5 climate models. *Journal of Climate*, **26**, 5846–5862, doi:10.1175/JCLI-D-12-00573.1.

1. Introduction

Projections of future climate change are usually inferred from numerical simulations made using complex computational climate models. These models attempt to approximate the behaviour of the Earth system as a deterministic dynamical system, represented by sets of discretised differential equations. However, there are many sources of uncertainty associated with projections of future climate. Statistical frameworks are required in order to quantify our uncertainty probabilistically, based on knowledge gained from climate models, observations, and our understanding of the Earth system.

But what is climate? The operational definition is a time average of weather, the most common being the 30-year average. However, if asked to describe the climate of Exeter in June, in terms of temperature, we would not respond with the sample mean temperature of early summer over the past 30 years, but with a range summarising the average temperature likely to be experienced during June in any given year. A definition that better reflects how we experience climate would be that climate is the probability distribution of weather (Stephenson et al., 2012; Rougier and Goldstein, 2014). By that definition, weather is a measurable property of the world, whereas climate can only be estimated using a statistical probability model. Our knowledge of the weather over the last 30 years is limited by our ability to observe it accurately and completely, this is *observation uncertainty*. However, our knowledge of the statistics of the distribution of weather over the past 30 years, i.e., the climate, is limited by *sampling uncertainty* and by our ability to define an appropriate probability model.

Uncertainty in climate projections is often broken down into three broad categories: initial condition, boundary condition, and model uncertainty. Non-linear interactions in complex dynamical systems, such as climate models, make them extremely sensitive to initial conditions. We can never know the precise state of the whole Earth system at any given time due to observation uncertainty. The slightest difference in initial conditions will result in the simulation of a very different sequence of future weather, even when forced by the same boundary conditions. The uncertainty about future climate due to initial condition uncertainty is often referred to as *unforced internal variability*.

Most global climate models include coupled atmosphere, ocean, sea ice and land surface components. Therefore boundary condition uncertainty is primarily due to human influences on the Earth system, in particular, the emission of greenhouse gases and changes to the land surface. However, future anthropogenic emissions and activities will be determined by complex socio-economic and political factors that cannot be predicted with certainty. This is usually referred to as *forcing uncertainty*, since the impacts of specific changes to the Earth system are often quantified by their effect on the balance of incoming and outgoing radiant energy. Fluctuations in the solar cycle, volcanic eruptions, and other unpredictable natural phenomena also contribute to uncertainty about future radiative forcing.

Model uncertainty reflects the fact that there are many different ways of constructing complex climate models. It is usually broken down into two parts. *Structural uncertainty* refers to the choices made in the process of constructing a climate model. Some physical processes may not be represented at all, if it is judged that their effect is likely to be small, or the computational complexity of including them would be too great. What type of grid is used to discretise the system, what horizontal and vertical resolution the model is run at, whether a simplified version of the equations of motion is implemented, and what numerical solvers are used can all affect the model outputs. The second type of model uncertainty arises due to the fact that many processes in the Earth system take place at scales that are smaller than the limited resolution of the models are able to resolve, e.g., cloud formation. The effects of these processes are instead incorporated by parameterising them, dependent on the large scale variables that are resolved by the models. The uncertainty about the best choice of parameters is referred to as *parameter uncertainty*.

The effect of internal variability can be quantified by repeatedly perturbing the initial conditions of a single climate model, for example, by starting the simulation from a different day in the same month (e.g., Deser et al., 2012b). Forcing uncertainty can be explored by running a single model with different scenarios of future greenhouse gas concentrations. In practice, projections from multiple future scenarios are rarely combined due to the difficulty of attaching prior probabilities to each scenario (Knutti et al., 2008). Parameter uncertainty can also be explored by perturbing the parameters of a single climate model, the resulting ensemble of outputs is usually called a perturbed physics ensemble (e.g., Forest et al., 2002; Harris et al., 2013; Knutti et al., 2006; Murphy et al., 2007; Sexton et al., 2012).

Parameter uncertainty, internal variability and forcing uncertainty can all be sampled using a single climate model. However, sampling structural uncertainty requires the consideration of multiple different models, i.e., a multi-model ensemble. Several large multi-model ensemble experiments have been coordinated under the supervision of the World Climate Research Programme's Coupled Model Intercomparison

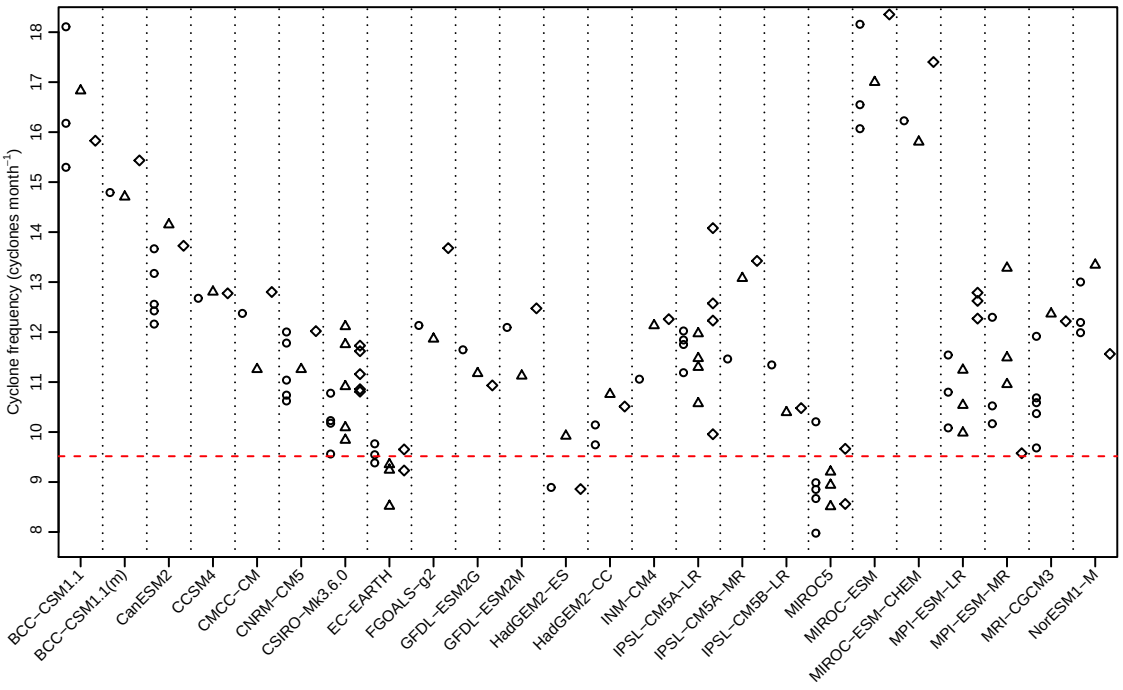

Figure 1.1.: 30-year time averaged winter (December-January-February) cyclone frequency in Exeter simulated by an ensemble of 24 climate models participating in the fifth coupled model intercomparison project. Each point represents one initial condition run, circles represent runs of the historical experiment (1976-2005), triangles and diamonds represent runs of the RCP4.5 and RCP8.5 future experiments (2070-2099) respectively. The red dashed line represents the observed cyclone frequency during the historical period, computed from the ERA-Interim reanalysis.

Project (CMIP). Each of the resulting databases contains multiple initial condition runs of several prescribed forcing scenarios by many climate models. The latest CMIP5 ensemble contains output from more than 40 climate models, submitted by more than 20 modelling centres around the world. Outputs from the CMIP5 climate change experiments are illustrated in Figure 1.1. Internal variability is represented by the spread of outcomes of the initial conditions runs from a particular experiment and model. Forcing uncertainty is represented by the differences between the outcomes of the two future experiments. Structural uncertainty is represented by the differences between the outputs of the different models.

A major problem is how to make inferences about future climate based on projections from multiple climate models? Climate science is unusual in its treatment of multiple models. Rather than treating them as incompatible and competing, each climate model is treated as a plausible representation of the climate system (Parker, 2006). This has resulted in two main paradigms for interpreting the relationship between climate models and the system they attempt to represent (Knutti et al., 2010a). The "truth plus error" paradigm assumes that the climates simulated by the models are a random sample from a distribution that is centred on the climate of the Earth

system. On the other hand, the "exchangeable" paradigm assumes that the climates simulated by the models and actualised in the Earth system are sampled from the same underlying distribution.

Neither of these interpretations is entirely satisfactory, for two main reasons. Climate models from different centres often share components, e.g., the same atmospheric circulation or ocean circulation model. Therefore, it is difficult to interpret the models as independent and a random sample. Also, climate models are fundamentally different from the system they represent. They are discretised approximations of the Earth system, and there are many processes present in the real system that are not represented in any model. Consequently, it is likely that shared errors may exist amongst the models. It is therefore difficult to imagine how the climate of the Earth system could be drawn from the same distribution. These and other difficulties associated with the interpretation of climate projections from multi-model ensembles, led the authors of a recent review to conclude that "quantitative methods to extract the relevant information and to synthesize it are urgently needed" (Knutti et al., 2010b).

The fact that computer simulations of physical systems are only approximations of the true systems means that even the best model we can conceive will never predict the behaviour of the true system exactly. This has been termed *model inadequacy* (Kennedy and O'Hagan, 2001). The uncertainty arising due to model inadequacy can be represented like any other uncertainty as a random quantity, usually called the *model discrepancy* (Craig et al., 2001). These concepts have previously been applied to the analysis of uncertainty in climate projections from perturbed physics ensembles (Sexton et al., 2012). However, the methodology has only recently been extended to the idea of an "ensemble discrepancy", i.e., the discrepancy between the climates simulated by an ensemble of models and the actual climate (Chandler, 2013; Rougier et al., 2013).

Despite the fact that each new generation of climate models is able to reproduce the recent climate with increasing accuracy, the spread of projections for key variables in the future has not decreased at the same rate (Knutti et al., 2008). This lack of improvement is a barrier to the exploitation of projections of future climate at regional and local scales. At smaller spatial scales, the effects of internal variability and model inadequacy both increase (Hawkins and Sutton, 2009). The models are unable to represent all of the small scale features of the Earth system and the advantage of smoothing by averaging over a large area is lost.

At the global level, the climate response simulated by a particular climate model, or combination of parameters in a model, is usually found to be independent of its ability to simulate the recent climate (Knutti et al., 2010b). However, at the local

level there are an increasing number of examples of "emergent constraints", i.e., relationships between the climate response and the recent climate simulated by a model, that emerge "consistently [...] from a wide range of detailed calculations (in particular, in this case, GCMs) rather than because of any physically direct calculation" (Ingram, 2010). Emergent relationships present promising opportunities to constrain projections of future climate using observations of recent climate. In a recent perspective, Collins et al. (2012) recommended that "that work is undertaken on both the theoretical underpinning and numerical implementation of the approach, so that it can be applied more widely".

1.1. Aims

The main objective of this thesis is the development of appropriate statistical frameworks in order to make credible inferences about long term climate change based on projections of future climate simulated by multiple climate models. Specific objectives are to develop frameworks that are able to

- separate the effect of internal variability simulated by the climate models and natural variability in the Earth system from other sources of uncertainty;
- reduce uncertainty about future climate change by exploiting emergent relationships to constrain future projections using observations of recent climate;
- account for the fact that projections from different climate models cannot be considered independent;
- account for the fact that *all* climate models are imperfect approximations of the Earth system.

In addition, there is a need to rigorously test such statistical frameworks, so that the resulting inferences can be considered credible.

1.2. Structure of this thesis

In Chapter 2, the difficulties associated with interpreting the outputs from ensembles of multiple climate models are discussed in more detail. Existing approaches to the problem of synthesising probabilistic projections of future climate change from multi-model ensembles are then reviewed. The developments contained in this thesis will be demonstrated by application to the estimation of possible changes in the frequency of extra-tropical cyclones over the North Atlantic, Europe and the Mediterranean basin. The chapter concludes with a brief discussion of how cyclone

activity is analysed in climate model output, and how it is likely to respond to future changes in climate.

In Chapter 3, the use of analysis of variance frameworks is explored for the analysis of multi-model ensemble climate change experiments. The emphasis in this chapter is on the contribution of internal variability to our uncertainty about the future climate response. Hypothesis tests are derived for evidence of climate model agreement, and of a non-zero climate response. We also address the question of whether or not existing multi-model ensembles are large enough to reliably detect future climate change. It is argued that a simple two-way analysis of variance framework can be used to estimate the future response of the actual climate, if all the climate models agree on the climate response.

In Chapter 4, the analysis of variance frameworks derived in Chapter 3 are extended in order to quantify structural uncertainty in addition to internal variability, when the models do not agree on the climate response. This allows the implementation of a cross-validation approach, in order to check that the statistical framework provides a good description of the variability present in the ensemble. The ensemble of climate models is reinterpreted using the Bayesian concept of exchangeability. Using this concept, the ensemble is systematically thinned in order to obtain a subset of the models that can be treated as a random sample.

In Chapter 5, the Bayesian hierarchical framework developed in Chapter 4 is extended to include the estimation of emergent constraints. It is argued that emergent constraints are properties of the differences between the expected climates simulated by the models, and do not apply to differences due to internal variability. It is shown that it is important to account for internal variability when estimating emergent constraints, however. The cross-validation approach is also extended to check the robustness of the estimated relationship.

In Chapter 6, the hierarchical framework developed in the previous chapters is extended to represent the uncertain relationship between the models and the Earth system as a discrepancy between the expected climate of an ensemble of climate models and the actual climate. Emergent relationships between the climate models are reinterpreted as constraints on the ensemble discrepancy. A simple method for estimating measurement error in the observations is proposed. The effects of both measurement error and sampling uncertainty are accounted for when combining observations with model output in order to constrain future projections. Projections of near surface temperature in the Arctic are compared with those from existing statistical frameworks that incorporate emergent constraints.

1.3. Original aspects of this thesis

It is common practice to include only one initial condition run from each climate model in a multi-model ensemble. One of the key methodological advances presented in this thesis is the inclusion of all available runs from each model. This allows differences between the preferred climates simulated by the models to be separated from differences due to unforced internal variability.

The distinction between model differences and internal variability is shown to be particularly important when estimating emergent constraints. To our knowledge, the methodology developed in this thesis is the first to allow for the effects of internal variability on the estimation of an emergent constraint from a multi-model ensemble.

The framework developed here builds on existing methods for representing model inadequacy. The methodology presented here is the first to explicitly interpret emergent relationships as constraints on the discrepancy between the climate responses simulated by the models, and the actual climate response.

A simple method is proposed for the estimation of measurement error in the observations from multiple reanalysis datasets. This allows the effects of measurement error and sampling uncertainty to be separated when combining climate model output with observations in order to constrain future projections.

2. Background

This chapter outlines the background to the developments contained in this thesis. It begins with a discussion of the issues associated with combining projections from multiple climate models, before reviewing the existing approaches to the problem. The methodology developed in the chapters that follow is illustrated by estimating the response to climate change of the frequency of extra-tropical cyclones over the North Atlantic and Europe. The response is estimated using the climate models participating in the latest Coupled Model Intercomparison Project, CMIP5 (Taylor et al., 2012). The current chapter concludes with a description of the CMIP5 ensemble, and how extra-tropical cyclone activity is analysed in climate model output.

2.1. What is climate?

In the introduction, climate was defined as the probability distribution of weather. The example of summer temperature in Exeter concerned only a single variable. More generally, climate can be thought of as a complex multivariate spatio-temporal process (Rougier and Goldstein, 2014). In this thesis, we restrict ourselves to the univariate case, and consider only a single climate variable. This thesis is concerned with long term climate projection, e.g., the climate at the end of 21st century. Long term climate projections are usually made in terms of 30-year averages of weather. However, if climate is the distribution of weather, then the distribution of 30-year averages of weather during any given period is also well defined. It is the statistics of this distribution, the distribution of 30-year averages of weather, that we aim to estimate. The expectation of this distribution will be of particular interest. Although we are primarily concerned with 30-year averages, the methodologies discussed in this chapter and developed in the chapters that follow may be applied to any time average of weather.

Before continuing, it is helpful to define some terminology. We loosely adopt the operational definition, and use the term *climate* to refer to any 30-year average of weather, either observed in the Earth system, or simulated by a climate model. The expected value of the distribution of 30-year averages of weather in the Earth system will be referred to as *the actual climate*. The equivalent quantity in a climate

model will be referred to as the *expected* or *preferred* climate of the model, or simply the climate of the model. The spread of the distribution of weather simulated by a climate model was defined in the introduction as the *internal variability* of the model. The spread of the distribution of the 30-year averages of weather in the Earth system will be referred to as *natural variability*. A particular 30-year average of weather in the Earth system, e.g., the average temperature in Exeter in June between 1970 and 1999, we will call *the climate that we experience*, or occasionally the *actualisation* of climate. A particular 30-year average of weather output from a climate model will be usually be referred to as a *run*, or occasionally a *realisation* of a climate model.

2.2. Interpreting multi-model ensembles

Multi-model ensembles are increasingly used by climate scientists to address the issue of model uncertainty in projections of future climate (Collins et al., 2012). Rather than viewing them as incompatible and competing, each climate model is considered as a plausible representation of the climate system (Parker, 2006). However, there are a number of issues that complicate the process of trying to synthesise projections from multiple climate models (Tebaldi and Knutti, 2007; Knutti et al., 2010b; Stephenson et al., 2012). These issues are briefly discussed below.

2.2.1. Ensemble design

In a perturbed physics ensemble (e.g., Collins et al., 2006a; Murphy et al., 2004; Stainforth et al., 2005), the experiment design is usually clear. Prior probability distributions for the input parameters are specified based on expert judgement. Parameter combinations are sampled systematically from those distributions so that the uncertainty about the inputs is properly represented. The climate model is run with one or more sets of initial conditions for each combination of input parameters. The outputs are then collated and used to form posterior distributions for the climate variables of interest (Sexton et al., 2012). Structural uncertainty cannot be explored since the experiment is limited to a single model, but the design is clear and easily interpreted from a statistical perspective.

Unfortunately, no such clarity is possible for multi-model ensembles. Ideally we would like to take the same approach as for perturbed physics ensembles - assign a prior distribution over the models, then sample climate models from that distribution. While defining the parameter space of a single model may be possible, defining a model space is problematic at best (Stainforth et al., 2007; Knutti et al., 2010b;

Stephenson et al., 2012). Without a well defined space to sample from, we cannot hope to design a systematic sample of climate models that fully explores the range of our structural uncertainty about how to model the Earth system. There are also good reasons to believe that the models we have do *not* represent a random sample, as we will see in the discussion that follows. Therefore, multi-model ensembles are often referred to as "ensembles of opportunity" (Allen and Ingram, 2002). One consequence of this convenience sampling is that the inter-model spread in the projections from multi-model ensembles has been found to depend strongly on which models are included (Knutti et al., 2008). This may also be partly due to the small number of models available.

These issues are exacerbated by the way in which multi-model ensembles are usually formed. Each modelling group will submit one or more initial condition runs of a set of prescribed forcing scenarios. No institute wants its model to be seen as performing poorly, so the best known parameter settings will be used for each model. As a result, the models included in the ensemble are a set of "best guesses", and are unlikely to span the full range of either parameter or structural uncertainty (Tebaldi and Knutti, 2007; Knutti et al., 2010b). The most prominent example of this is climate sensitivity (the change in global mean temperature in response to a doubling of atmospheric CO_2). The CMIP3 models (Meehl et al., 2007) simulate a range of climate sensitivity of approximately 2.0-4.5 K (Meehl and Coauthors, 2007, Box 10.2). Perturbing the parameters of a single model may yield much larger values of climate sensitivity (Murphy et al., 2004; Stainforth et al., 2005). However, a multi-model ensemble may still be more effective for local changes, or where the response is dominated by processes that are less well understood and represented very differently in different climate models.

2.2.2. Model dependence

It was noted above that the models included in a multi-model ensemble are unlikely to represent a random sample of our structural uncertainty about the climate system. Although models may be developed by different scientists working in different centres around the world, they may not be as dissimilar as might be hoped. At a practical level, they share the same basic problem of trying to approximate a continuous system by a discretised set of equations. Similar numerical methods are used to solve those equations. The current state of the art in computing technology limits the models to resolving similar horizontal and vertical resolutions (Stainforth et al., 2007). At an intellectual level, the models are based on the same equations and the same knowledge gathered from the same sources. Methods that perform well are published, and then adopted and included in other models.

Due to the cost and complexity of developing a fully coupled climate model, developers also share parameterisations or even entire components so that several models may utilise the same atmosphere or ocean components. In the CMIP3 (Meehl et al., 2007) and CMIP5 (Taylor et al., 2012) ensembles, several modelling centres submitted runs from multiple models. In practice, those different models might involve swapping one ocean model for another, adding an atmospheric chemistry module, or simply running the same model at a different resolution. Models that share components will not contribute as much information about the range of possible structural uncertainty as models constructed from unique components. In that sense, many of the climate models included in a multi-model ensemble cannot be considered to be independent of one another (Masson and Knutti, 2011).

In statistical terms, the outputs of the models are likely to be correlated with one another. This has been demonstrated in a number of studies by comparing model outputs to observations, and to each other (Jun et al., 2008; Knutti et al., 2010b). Hierarchical cluster analysis has been used to show that the outputs of climate models from the same centre, or which share components, are more similar than those from models developed "independently" (Masson and Knutti, 2011; Knutti et al., 2013). An attempt to estimate the effective number of models yielded an approximate value of 7.5-9.0 models for the 24 model CMIP3 ensemble (Pennell and Reichler, 2011). Unless the correlation between models is accounted for, any probabilistic estimate of the climate response is likely to be overconfident.

Components or methods common to multiple climate models also admit the possibility of shared errors or common biases. There is empirical evidence that there may be biases common to *all* climate models (Annan and Hargreaves, 2010; Knutti et al., 2010b). In addition to the issues outlined above, there are also processes that are not included in any contemporary climate model. They are excluded either due to lack of understanding, the belief that the effect of their inclusion would be small, or because the computational burden of including them would be too great. However, their exclusion represents a structural deficiency common to all the models. The effects of these missing processes are likely to manifest themselves as shared biases.

2.2.3. Model evaluation

Quantitative evaluation of climate model performance for century scale climate projection is a fundamentally different problem than evaluating numerical weather prediction or seasonal climate prediction models. In weather forecasting, a constant supply of paired forecasts and observations is available that can be used to assess the predictive ability of a model. From a statistical perspective, uncertainty in weather forecasting conforms to the frequentist point of view and can be quantified

from a long series of repeated events. For century scale climate projection, no similar predictive confirmation of the models is possible, so confidence in projections must come from other sources (Tebaldi and Knutti, 2007; Knutti et al., 2010b). In the absence of repeated trials, confidence in climate projections is necessarily based on our beliefs about the models. Therefore probabilities associated with projections of future climates have a subjective Bayesian interpretation (Weigel et al., 2010; Stephenson et al., 2012). The problem of predictive confirmation cannot be solved by simply waiting for data to become available. By the time data is available, the models used to make the projections will long since have been retired (Smith, 2002). In addition, long term climate projections are conditional on forcing scenarios that are unlikely to accurately reflect anthropogenic emissions of greenhouse gases over the next century (Allen et al., 2013). So it would not be a fair comparison.

How then should we go about quantifying climate model performance? The obvious answer is by their ability to reproduce aspects of the recent climate according to some metric. There have been many attempts to define metrics of model performance, usually based on biases compared to one or more observed climate means or trends (e.g., Lambert and Boer, 2001; Murphy et al., 2004; Gleckler et al., 2008; Reichler and Kim, 2008). Unfortunately, no one model performs best for all variables or in all regions (Lambert and Boer, 2001; Jun et al., 2008). This immediately raises the question of which variables and processes are most important for assessing model performance, and what properties of those elements (mean, trend, seasonal cycle, etc.) should any combined metric be based on? It has been argued that since the climate system is ultimately driven by incoming short-wave radiation from the sun, then model performance in representing radiative fluxes should be prioritised (Huber et al., 2011). However, even restricting ourselves to considering one set of variables, there are likely to be dependencies between those variables and their properties of interest. So it is not even clear how performance measures should be combined.

Suppose that we could define a metric that we were confident represented model performance in reproducing recent climate, how then should we interpret such a metric? If a model is unable to adequately reproduce recent climate, then we should certainly question its ability to reliably simulate future climate (Oreskes et al., 1994). However, the ability to reproduce recent climate, does not guarantee any skill in simulating long term climate change. Projecting climate change is inherently an extrapolation problem (Stainforth et al., 2007). We are trying to predict a state never before seen in the instrumental record, although analogues may exist in the paleoclimate record. Parameterisations that work well today may be inadequate in a changed climate, and processes that are not currently represented well (or at all) in the models may become important. Therefore, good performance in reproducing recent climate (or recent climate change) is not necessarily a reliable indicator of

performance for simulating future climate change.

2.2.4. Model tuning

The tuning of model parameters further complicates the process of climate model evaluation. Due to the limited amount of data available, the same observations against which the models are evaluated may already have been used in the tuning process. The possibility then exists that a particular model may appear to perform well simply because it has already been tuned to reproduce a particular set of observations.

The risk may not be so great as it seems, however. Due to the complexity of global climate models, they are usually calibrated piecewise based on a mixture of performance metrics and expert judgement (Knutti et al., 2010b). The number of runs required to optimise performance over a large number of input parameters simultaneously would be prohibitive. The computational expense would be too great. In practice, the empirical method is found to be quite effective when compared to the results of a large perturbed physics ensembles, at least when combined performance is measured over a number of variables (Sanderson et al., 2008).

Different combinations of parameter values may result in similar performance (Stainforth et al., 2005). Given the difficulty in defining meaningful performance metrics, it is conceivable that particular parameter combinations may result in apparently good performance for the wrong reasons. If biases in one process are compensating for biases in another (e.g., Knutti et al., 2002), then the response to climate change may be poorly simulated. Piecewise calibration means that such effects may go unnoticed. Such dependencies should be discovered as part of any well designed multi-parameter calibration scheme. However, expert judgement would still be required to select a "best" combination until additional evidence or increased understanding allowed the parameters to be constrained further.

2.2.5. Reasons to trust climate models

Given the problems outlined above, in particular the difficulties in evaluating model performance, why should we have confidence in climate models? Knutti (2008a) set out a number of subjective reasons for trust in climate models, which are briefly reviewed here. Most importantly, models are built on sound physical principles, e.g., the equations of motion, conservation of energy, momentum and mass. Models are able to reproduce many aspects of recent climate reasonably well (Räisänen, 2007; Randall et al., 2007; Gleckler et al., 2008) and performance in doing so has continued

to improve with each new generation (Reichler and Kim, 2008). Climate models are also able to reproduce large scale trends in recent climate (Barnett et al., 2005; Hegerl et al., 2007; Knutti, 2008b). Although there is some possible discrepancy with the hiatus in warming observed since the mid 1990s (Fyfe et al., 2013). Our understanding of climate feedback processes is also increasing (Bony et al., 2006; Soden and Held, 2006), which should help minimise the likelihood of compensating biases in models. Paleoclimate data also provides a promising source of additional data against which to evaluate models (Jansen et al., 2007). For the first time, the latest Coupled Model Intercomparison Project includes a set of coordinated paleoclimate experiments (Taylor et al., 2012). This will allow model performance to be compared under very different conditions to those in the recent observational record. So despite the many difficulties, there are also some important reasons for confidence in climate models.

2.3. Weighting climate models

While the prevailing view might be that all climate models are plausible representations of the climate system (Parker, 2006), it is natural to consider that some models might be more plausible than others. When estimating future climate change, we might wish to attach additional weight to projections from models that show better performance. The simplest example of model weighting is that older models are rarely included, if they lack developments considered mandatory in contemporary models. Discarding a model completely is equivalent to assigning it zero weight. There are two main problems associated with model weighting. As discussed earlier in this chapter, it is difficult to define a single metric of model performance. Also, any metric that is defined is necessarily based on the ability of a model to reproduce past climate. However, it has been shown that if the weights do not reflect the true ability of the model to respond to climate change, then the resulting projections are likely to be less accurate than if equal weight was given to all the models (Weigel et al., 2010).

Despite the associated difficulties, a number of approaches to weighting projections from climate models have been proposed. In seasonal climate forecasting it is common to regress time series of observations on time series of model outputs to obtain an "optimal" set of model weights (e.g., Krishnamurti et al., 2000; Kharin and Zwiers, 2002; DelSole, 2007; Peña and van den Dool, 2008). These methods have also been adapted to long term climate change (e.g., Greene et al., 2006; Bishop and Abramowitz, 2013). Other studies have weighted models based on their biases compared to observations of recent climate (e.g., Giorgi and Mearns, 2002; Tebaldi et al., 2005; Smith et al., 2009). The mean squared error compared to observa-

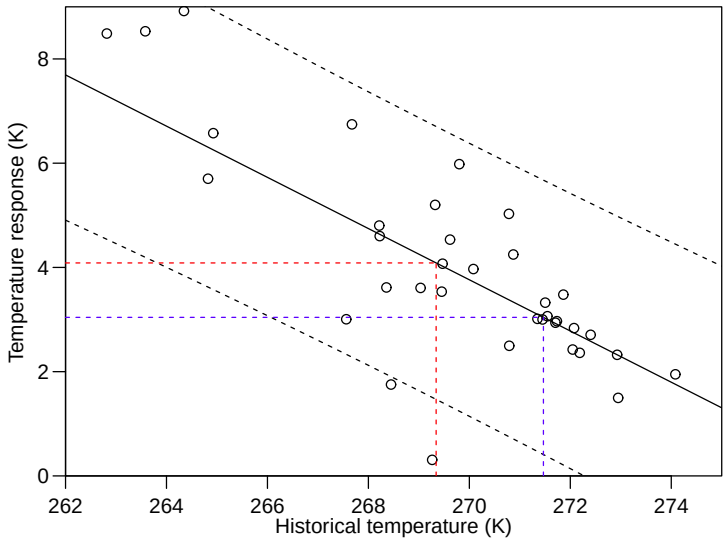

Figure 2.1.: Winter (December-January-February) mean near surface (2m) air temperature change in 2069-2099 under the RCP4.5 forcing scenario compared to 1975-2005 temperature in the Bering Sea (179E,59N), simulated by an ensemble of 37 climate models participating in the fifth coupled model inter-comparison project. Each point represents the mean of all available initial condition runs by one model. The solid black line represents the linear regression of the temperature change on the historical temperature. The black dotted lines are a 95% prediction interval for the response of a new model. The dashed red lines represent the mean historical temperature and temperature response of the models. The dashed blue line represents the observed historical temperature estimated from the ERA-Interim reanalysis, and the response projected by linear regression.

tions of one or more variables has also been used (e.g., Connolley and Bracegirdle, 2007; Pierce et al., 2009; Reifen and Toumi, 2009). Some methods consider combinations of application specific metrics (e.g., Waugh and Eyring, 2008; Christensen et al., 2010). While others take widely varying approaches to pattern scaling (e.g., Shiogama et al., 2011; Watterson and Whetton, 2011).

If it could be demonstrated that model performance in simulating recent climate was well correlated with performance in simulating the projected response to climate change, then performance based weighting might be justified. Until recently, it was thought that such correlations were rare (Whetton et al., 2007; Jun et al., 2008; Knutti et al., 2010b). At the global scale, that may be the case. However, at local and regional scales there are an increasing number of examples of "emergent constraints" on the climate response (e.g., Hall and Qu, 2006; Boé et al., 2009; Bracegirdle and Stephenson, 2012). These present promising opportunities to constrain projections of future climate, and are discussed in detail below.

2.4. Emergent constraints

The term "emergent constraint" was first used in climate science by Allen and Ingram (2002) to describe "constraints relating past to future greenhouse warming that seem to hold across all available climate models". More generally, an emergent constraint might be defined as a relationship between the future climate change in one variable, and the state of a related variable in the past, that is robust across many climate models. For example, Bracegirdle and Stephenson (2012) found that the near-surface (2m) temperature change simulated by climate models in the Arctic is negatively correlated with the present day temperature simulated by the same models (Figure 2.1). They used simple linear regression to estimate the relationship between the model outputs, and then compared with recent observations of Arctic temperature to obtain a constrained estimate of future warming.

Several other examples of emergent constraints have been identified in the Arctic (Hall and Qu, 2006; Boé et al., 2009; Räisänen et al., 2010; Mahlstein and Knutti, 2011). The strength of the present day seasonal cycle in snow albedo feedback has also been used to constrain the snow albedo feedback on future climate change (Hall and Qu, 2006). Chemistry climate models simulating lower peak levels of stratospheric inorganic chlorine between 2000-2010 have been shown to simulate earlier ozone hole recovery (Eyring et al., 2007; Karpechko et al., 2013). Emergent constraints have also been identified in the carbon cycle (Cox et al., 2013; Wenzel et al., 2013). An extensive examination of radiative fluxes simulated by climate models has been used to constrain estimates of climate sensitivity (Huber et al., 2011). Additional examples have shown relationships between simulation of stratospheric ozone and Southern hemisphere tropospheric circulation (Son et al., 2010), regional temperature and precipitation (Schaller et al., 2011), and present day relative humidity and climate sensitivity (Fasullo and Trenberth, 2012). One study has also linked the sensitivity of extreme values of tropical precipitation to monthly mean temperature anomalies with their sensitivity to climate change (O'Gorman, 2012). Several studies have found emergent constraints linked to the present day variability of climate variables (Hall and Qu, 2006; Cox et al., 2013; O'Gorman, 2012). If a model underestimates the strength of the response to seasonal or multi-annual forcing, then it is also likely to underestimate the response to long term climate change, since many of the same mechanisms will be involved on both time scales.

Ingram (2010) defined an emergent constraint as "something we believe because we consistently get the same answer from a wide range of detailed calculations (in particular, in this case, GCMs) rather than because of any physically direct calculation". However, the possibility remains that the relationship may be an artefact due to some common deficiency amongst the models, or the small number of models

available. Climate scientists have been quick to address this issue. Plausible explanations based on detailed understanding of climate processes have accompanied most of the emergent relationships outlined above. For example, the Arctic temperature constraint described by Bracegirdle and Stephenson (2012) was strongest close to the sea ice edge, where it was linked to systematic biases in the simulation of sea ice extent. There is still a danger inherent in such post-hoc interpretation, however the reassurance is welcome.

2.5. Existing approaches to synthesising climate projections from multi-model ensembles

We have reviewed the main problems associated with interpreting projections from multi-model ensembles. With those issues in mind, we now go on to discuss existing approaches to synthesising projections from multi-model ensemble climate change experiments. First we describe the most commonly used heuristic method of combining projections from multiple models, the multi-model mean. Based on the preceding discussion, several simple criteria are suggested against which the credibility of the more formal methods can be judged. Tebaldi and Knutti (2007) presented a chronological review of the methodologies proposed at the time. Here we take a different approach and break down those methods, and several novel methods proposed in the interim, into the following categories

1. Heuristic averages - the multi-model mean
2. "Truth plus error" methods
3. "Exchangeable" methods
4. Discrepancy methods
5. Ensemble regression
6. Constant relationship methods

Other classifications are possible, but these help to distinguish the projection methods according to their underlying assumptions.

In order to facilitate comparison between the various approaches, the statistical formulations will be given in a common notation where possible. Symbols X_m and x_m denote a random variable representing the output of climate model m and a realisation of that model output, respectively. Unless otherwise stated, X_m is assumed to represent the output from a single initial condition run. Similarly, Y and y denote the climate of the Earth system, usually the actual climate, i.e., the

expectation of the distribution of time averaged weather. Finally, Z and z represent observations of the Earth system. Additional subscripts are used to denote the time periods / forcing scenarios considered, H denotes the historical period, F denotes a future scenario, and in some cases R is used to denote a response (the difference between a future and a historical scenario). Standard statistical notation is used for variances which are denoted by σ^2, and precisions (the reciprocal of variance) which are denoted by τ.

2.5.1. Heuristic averages - multi-model mean

Estimates of the future climate change response from multi-model ensembles are often based on a multi-model mean

$$\bar{X}_R = \frac{1}{M} \sum_{m=1}^{M} X_{Rm}$$

where X_{Rm} represents the climate response simulated by model m, and M is the number of models (e.g., Collins et al., 2013; Meehl and Coauthors, 2007). There is empirical evidence from seasonal climate forecasting that the multi-model mean, will often outperform any single model on a range of measures of predictive ability (Doblas-Reyes et al., 2003; Palmer et al., 2004; Hagedorn et al., 2005). It has also been shown that the multi-model mean tends to outperform most individual models in terms of their mean-squared-error in reproducing historical climate (Lambert and Boer, 2001; Gleckler et al., 2008; Knutti et al., 2010b). This result has proved to be robust across several generations of climate models (Reichler and Kim, 2008; Flato et al., 2013). The multi-model mean may not outperform all single models for every variable. However, it is found to perform consistently well when tested over multiple variables, since no single model performs best for all variables, or in all regions (Hagedorn et al., 2005).

2.5.2. What makes a credible representation of climate and climate models?

Before reviewing more formal methods, it is useful to consider some simple criteria against which the credibility of the proposed approaches can be judged. In Section 2.2, it was argued that climate models could not be considered independent due to shared components and parameterisations. There are also processes that are not represented in any climate model and may result in biases or errors shared by all models. Two empirical results are often cited as evidence that climate models are not independent, and may contain shared errors. First, model biases relative

to historical observations are often found to be correlated (Jun et al., 2008; Knutti et al., 2010b). Second, the mean squared error of the multi-model mean relative to historical observations does not converge to zero as more models are added to the ensemble (Knutti et al., 2010b; Annan and Hargreaves, 2010). These empirical results suggest that in order to be considered credible, a statistical framework for making inferences about the actual climate from an ensemble of climate models should predict the following properties

$$\operatorname{cov}\left(X_{Hi} - Y_H \ , \ X_{Hj} - Y_H\right) \neq 0 \tag{2.1a}$$

$$\lim_{M \to \infty} \mathrm{E}\left(\left(\bar{X}_H - Y_H\right)^2\right) \neq 0 \tag{2.1b}$$

where X_{Hm} is the historical climate of model m. We write Y_H for the actual climate, or the climate that we experienced, rather than Z_H for observations since measurement error is usually assumed to be small. However, if the measurement error were not negligible, then this might help explain both phenomenon.

The motivation for synthesising projections from ensembles of climate models is to explore our structural uncertainty about how to model the climate system. Therefore, we should expect our uncertainty about the climate response in the Earth system Y_R to span the range of climate responses simulated by the models (Lopez et al., 2006; Tebaldi and Sansó, 2009; Rougier et al., 2013), so that

$$\operatorname{var}\left(Y_R\right) \geq \operatorname{var}\left(X_{Rm}\right) \tag{2.2}$$

These three simple criteria provide a useful reference as we review the various approaches that have been proposed for synthesising projections from ensembles of climate models.

2.5.3. The "truth plus error" approach

In the third assessment report of the Intergovernmental Panel on Climate Change, it was suggested that the climate or climate response simulated by model m could be represented as

$$X_m = Y + R_m + \varepsilon_m \tag{2.3}$$

where R_m is the departure of model m from the actual climate Y due to model error, and ε_m is the departure due to internal variability (Cubasch et al., 2001, Section 9.2). This representation has become known as the "truth plus error" interpretation of climate and climate models (Knutti et al., 2010a). If the model errors R_m are independent of one another, and likewise the departures due to internal variability ε, then the multi-model mean will converge to the actual climate Y in a large ensemble

(Cubasch et al., 2001; Tebaldi and Knutti, 2007; Knutti et al., 2010b)

$$\bar{X} \to Y \quad \text{as} \quad M \to \infty$$

Therefore, most of the projections presented in the third assessment report, and each subsequent report, have been based on the means of multi-model ensembles (Houghton et al., 2001; Solomon et al., 2007; Stocker et al., 2013).

Reliability ensemble averaging

Giorgi and Mearns (2002) proposed a method of weighting projections from multiple models called "reliability ensemble averaging" based on two performance measures. They suggested that climate models that simulate small biases compared to observations of recent climate, *and* that simulate a climate response that agrees with the consensus of the other models, should be seen as more reliable. These two measures were referred to as "bias" and "convergence" criteria, respectively. Model projections were combined as a weighted average to give an estimate of the actual climate response

$$y_R = \frac{1}{\sum_m w_m} \sum_{m=1}^{M} w_m x_{Rm} \tag{2.4}$$

where y_R is the actual climate resopnse and x_{Rm} is the climate response simulated by climate model m. The weights w_m are defined so that only a model that performs well on both the bias and convergence criteria will receive a large weight

$$w_m = \left\{ \left[\frac{\epsilon}{|x_{Hm} - z_H|} \right]^a \left[\frac{\epsilon}{|x_{Rm} - y_R|} \right]^b \right\}^{\frac{1}{a \times b}} \tag{2.5}$$

where x_{Hm} is the historical (recent) climate simulated by model m, and z_H is the observed climate. The first component of the weights depends on the model bias $x_{Hm} - z_H$, while the second component measures the convergence $x_{Rm} - y_R$ of the model response to the estimated response. The constants a and b allow the relative importance of the two criteria to be adjusted. Only one initial condition run is included from each model, so the model departures will also include a contribution due to internal variability. The parameter ϵ is an estimate of the internal variability, and ensures that models are not discounted unless their bias and deviation from the consensus are large compared to the unforced variability. Since the consensus is defined relative to the weighted estimate y_R, an iterative procedure is required to fit the model weights. Nychka and Tebaldi (2003) later showed that the weighted estimate of the climate response given by Equations 2.4 and 2.5, is equivalent to the median of the model responses x_{Rm}, weighted only by their bias from the observations. If the model responses are distributed symmetrically about the actual

climate response, then the median is a more robust estimate of the actual climate response (Nychka and Tebaldi, 2003).

A probabilistic interpretation of reliability ensemble averaging

The reliability ensemble averaging estimate is a heuristic estimate with no formal distributional assumptions, therefore it only provides a point estimate. A probabilistic interpretation of the reliability ensemble averaging approach was developed by Tebaldi et al. (2004, 2005). The basic structure of the framework they proposed is

$$Z_H = Y_H + W_H \tag{2.6a}$$
$$X_{Hm} = Y_H + R_{Hm} \tag{2.6b}$$
$$X_{Fm} = Y_F + R_{Fm} \tag{2.6c}$$

where X_{Fm} represents the future climate simulated by model m, and Y_H is the actual historical climate. The response of the actual climate is estimated by $Y_R = Y_F - Y_H$ The term W_H represents the departure of the observed climate Z_H from the actual climate due to natural variability. The model departures R_{Hm} and R_{Fm} include contributions from both model uncertainty and internal variability, since only one initial condition run is included from each model. The departures were assumed to be independent (between models) but *not* identically distributed

$$R_{Hm} \sim N\left(0\ ,\ \tau_m^{-1}\right) \tag{2.7a}$$
$$R_{Fm} \sim N\left(\lambda R_{Hm}\ ,\ (\theta\tau_m)^{-1}\right) \tag{2.7b}$$

The model specific precisions τ_m represent the tendency of each model to deviate from the actual climate. The parameter θ allows for the possibility that the models will tend to deviate more strongly from the actual climate in the future. The future departure of each model R_{Fm} is conditioned on its historical departure R_{Hm}. The parameter λ controls the strength of the correlation. Note that if $\lambda \neq 0$ or 1, then the response $X_{Fm} - X_{Hm}$ simulated by model m will be correlated with its historical climate, i.e., λ represents an emergent constraint. This is an interesting choice since such relationships were thought to be rare at the time (Giorgi and Mearns, 2002).

The framework proposed by Tebaldi et al. (2005) was formulated from a Bayesian perspective. Prior probability distributions were specified for all of the unknown parameters. Gamma priors were specified for the model specific precisions τ_m. The Normal-Gamma mixture formulation is equivalent to assuming that the model departures R_{Hm} and R_{Fm} are t distributed (Gelman et al., 2014). The heavy tails of the t distribution make it a common choice for making estimates robust against

outlying data points, similar to reliability ensemble averaging (Nychka and Tebaldi, 2003). In fact, the expected values of the posterior distributions of the actual historical and future climate, Y_H and Y_F, can be written as weighted averages similar to Equation 2.4, but weighted by the model specific precisions τ_m (Tebaldi et al., 2005). The posterior expectations of the precisions τ_m (conditional on the other parameters) are given by (Tebaldi et al., 2005, Equation 12)

$$\mathrm{E}\left(\tau_m \mid \ldots\right) = \frac{a+1}{b + \frac{1}{2}\left\{\left(x_{Hm} - y_H\right)^2 + \theta\left[x_{Fm} - y_F - \lambda\left(x_{Hm} - y_H\right)\right]^2\right\}} \tag{2.8}$$

where a and b are prior constants, chosen to be small compared to the other terms so that their influence on the posterior is minimised. When $\lambda = 1$, equivalent to the response of model m being independent of its historical climate, these are essentially the reliability ensemble averaging weights in Equation 2.5 (Tebaldi et al., 2005).

In the original formulation, identical but separate priors were specified for each of the model specific precisions τ_m, i.e., they were assumed to be systematically different for each model. Instead, Smith et al. (2009) modelled the τ_m as arising from a common distribution whose parameters were also estimated. This has the effect of constraining the weights to be more similar to each other, reducing the tendency for a small number of models to dominate the projections (Tebaldi and Knutti, 2007). A cross-validation step was also introduced in order to check the statistical assumptions (Smith et al., 2009). The final innovation proposed by Smith et al. (2009) was the extension to simultaneous estimation over a set of predefined geographical regions. The model specific precisions will better reflect the true performance of each model in this formulation, since they are evaluated over multiple projections (Tebaldi and Knutti, 2007).

Modelling the spatial structure of the climate response

Furrer et al. (2007b,a) took a different approach and proposed a joint framework for the climate response at all grid points. They did this by splitting the response simulated by each model into a large scale signal, and small scale noise process over space

$$X_{Rm} = \mathbf{X}\boldsymbol{\beta}_m + \boldsymbol{\varepsilon}_m \tag{2.9}$$

where X_{Rm} is the vector made up of the climate response simulated by model m at each grid point. The design matrix $\mathbf{X}$ is a matrix of spatial basis functions, and $\boldsymbol{\beta}_m$ is a vector of regression coefficients associated with model m. Together these represent the large scale climate change signal simulated by model m. The vector $\boldsymbol{\varepsilon}_m$ represents the small scale noise in the simulated climate response, due to local effects, internal variability etc.. The noise was modelled as a zero mean

random process with a covariance structure that depended only on the great circle distance between two grid points. Observations of the recent climate $\mathbf{z}$ are included in the design matrix $\mathbf{X}$ in order to partially constrain the large scale structure of the response to reflect the spatial distribution of the historical climate.

The crucial assumption is that the regression coefficients $\boldsymbol{\beta}_m$, while different for each model, are treated as random quantities that are assumed to arise from a common multivariate normal distribution, the expected value of which corresponds to the actual climate

$$\boldsymbol{\beta}_m \overset{iid}{\sim} MVN\left(\boldsymbol{\beta}_Y \ , \ \boldsymbol{\Sigma}\right) \tag{2.10}$$

where $\boldsymbol{\beta}_Y$ is the expected value of the regression coefficients and $\boldsymbol{\Sigma}$ is a covariance matrix. So Equation 2.9 can be rewritten in the "truth plus error" form of Equation 2.3 by letting $Y = \mathbf{X}\boldsymbol{\beta}_Y$ and

$$R_m = \mathbf{X}\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_Y\right) \overset{iid}{\sim} MVN\left(\mathbf{0} \ , \ \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T\right) \tag{2.11}$$

Since the model departures R_m are assumed to be identically distributed, all models are treated equally in this framework.

Generalising the "truth plus error" approach

Unfortunately none of the approaches outlined above satisfy our simple credibility criteria. For the framework proposed by Tebaldi et al. (2005) we have

$$\begin{aligned} \operatorname{cov}\left(X_{Hi} - Y_H \ , \ X_{Hj} - Y_H\right) &= 0 \\ \mathrm{E}\left(\left(\bar{X}_H - Y_H\right)^2\right) &= \frac{1}{M^2}\sum_{m=1}^{M} \operatorname{var}\left(R_{Hm}\right) \end{aligned}$$

so the biases are expected to be uncorrelated and the mean squared error of the multi-model mean are expected to converge to zero as the ensemble size increases. Furrer et al. (2007b) only specify assumptions about the climate response, however the same formulation could be easily applied to the historical climate. But since it has the same basic "truth plus error" form of Equation 2.3 and the model departures R_m are assumed to be independent with mean zero, the conclusion is the same. Further, by treating the actual climate (or climate response) Y as the central tendency of a sample of independent climate models, the estimated uncertainty for Y will tend to be inversely proportional to the number of models, i.e., $\operatorname{var}\left(Y_R\right) \ll \operatorname{var}\left(X_{Rm}\right)$ (Lopez et al., 2006; Tebaldi and Knutti, 2007; Knutti et al., 2010b).

Tebaldi and Sansó (2009) address several of these issues in a further extension of the methodology developed by Tebaldi et al. (2005) and Smith et al. (2009). For a single variable, the basic structure of the proposed framework is

$$Z = Y + W \tag{2.12a}$$

$$X_m = Y + B + R_m + \varepsilon_m \tag{2.12b}$$

where X_m, Y and Z are vectors of random variables representing time series of the climate simulated by model m, the actual climate and the observed climate respectively. Once again, W represents departures due to natural variability. A piecewise linear trend over time is specified for the actual climate Y (not shown). The models are all assumed to simulate the same (correct) trend. The climate response Y_R is estimated by the mean difference in climate Y between two time periods. The model departures R_m are now assumed to be independent *and* identically distributed with mean zero and common variance $\mathrm{var}(R_m)$, so no model weighting takes place in this framework. The model departures are assumed to be constant over time. The vector ε_m represents the departures of model m due to internal variability. For each model, ε_m is modelled as mean zero but with a unique variance $\mathrm{var}(\varepsilon_m)$, representing the magnitude of the internal variability as simulated by model m. The new term B represents any bias common to all the models. Like the R_m, it is also assumed to be constant over time, but has no parameters associated with it. Due to the inclusion of B, the multi-model mean will not converge to the actual climate as the ensemble size increases

$$\mathrm{E}\left(\left(\bar{X} - Y\right)^2\right) = B^2 + \frac{1}{M}\,\mathrm{var}(R_m) + \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{var}(\varepsilon_m)$$

although the individual model biases are still uncorrelated in this formulation, since B is treated as constant. The authors acknowledge that the posterior distribution of the climate response derived from Y will not span the full range of responses simulated by the models. They suggest that the posterior predictive distribution of a new model X^*_m will better represent the uncertainty about the response of the actual climate (Tebaldi and Sansó, 2009). While this would span the range of responses simulated by the models, it is an ad-hoc interpretation and would benefit from additional statistical formalism.

2.5.4. The "exchangeable" paradigm

Some of the earliest studies to derive probabilistic projections from an ensemble of climate models used the empirical distribution defined by the model outputs to estimate the probabilities of events of interest (e.g., Räisänen and Palmer, 2001).

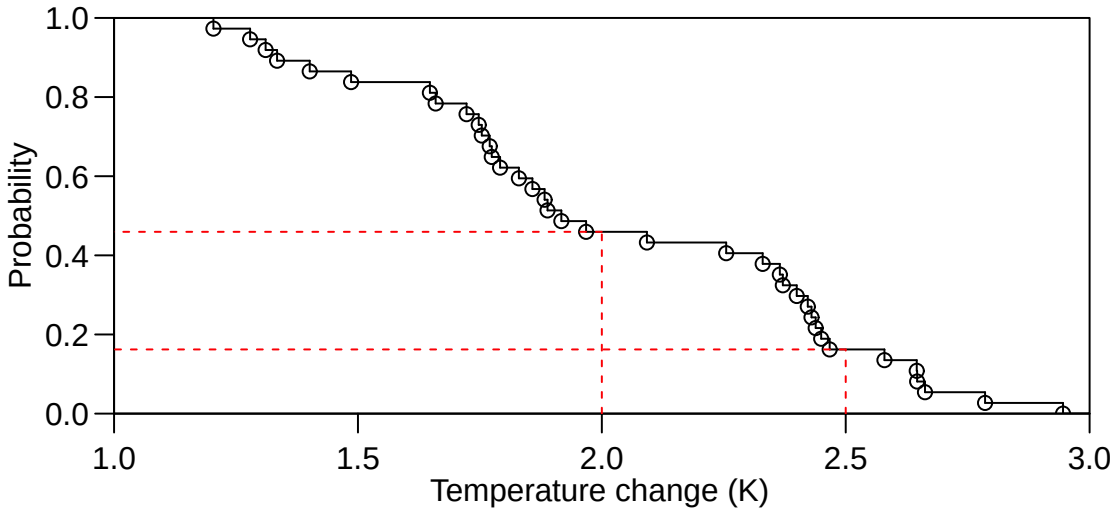

Figure 2.2.: The empirical cumulative distribution function of the change in global mean temperature in 2070-2099 compared to 1970-1999 under the RCP4.5 forcing scenario, computed from an ensemble consisting of one run from each of 37 climate models participating in the fifth coupled model inter-comparison project. Each point represents the climate response simulated by one model. The dashed red lines indicate the estimated probability of the change in global mean temperature exceeding 2.0K or 2.5K.

For instance, the probability of the global mean temperature change in a particular future scenario exceeding 2K would be estimated by the proportion of models simulating a response above that threshold (Figure 2.2). This was the approach adopted by Giorgi and Mearns (2003), except the model outcomes were weighted using the weights defined by reliability ensemble averaging (Giorgi and Mearns, 2002) so that

$$\Pr\left(Y_R \geq y_R\right) = \frac{\sum_{m=1}^{M} w_m \, \mathrm{I}\left(X_{Rm} \geq y_R\right)}{\sum_{m=1}^{M} w_m} \tag{2.13}$$

where the w_m are the weights and I is the indicator function that takes the value 1 when its argument is true, and 0 otherwise. If the model weights are all equal, then this reduces to the basic method of Räisänen and Palmer (2001). The underlying assumption is that the either the actual climate, or the climate response that we will experience, is drawn from the distribution defined by the responses of the models, and so the model spread quantifies our uncertainty about the actual response. This is quite different from the "truth plus error" approach, where the models are assumed to be centred on the actual climate response. Instead, the actual response may lie anywhere within the range of responses simulated by the models.

Interest in this applying this approach to climate projection was revived sometime later by Annan and Hargreaves (2010, 2011) who suggested the assumption that the actual climate is "exchangeable" with, or "statistically indistinguishable" from, the climates simulated by the models, i.e., drawn from the same distribution (the

Bayesian concept of exchangeability is discussed in more detail Chapter 4). They argued that this approach addressed a number of the issues with the "truth plus error" interpretation.

$$\begin{aligned} \operatorname{var}(Y) &= \operatorname{var}(X_m) \\ \mathrm{E}\left(\left(\bar{X} - Y\right)^2\right) &= \frac{1}{M}\operatorname{var}(X_m) + \operatorname{var}(Y) = \frac{1}{M}\operatorname{var}(X_m) + \operatorname{var}(X_m) \\ \operatorname{cov}(X_i - Y, X_j - Y) &= \operatorname{var}(Y) = \operatorname{var}(X_m) \end{aligned}$$

Since the actual climate Y is assumed to be drawn from the same distribution as the climates simulated by the models X_m, it has the same variance. So our uncertainty about the climate response of the Earth system will span the range of responses simulated by the models. The mean squared error of the multi-model mean is limited by that same variance, and the model biases are expected to be correlated. So the "exchangeable" paradigm satisfies all three of our simple criteria for a credible representation of the relationship between the ensemble and the Earth system.

2.5.5. Discrepancy methods

The idea of a discrepancy between the climate simulated by a model and the actual climate has previously been included in the analysis of perturbed physics ensembles (Murphy et al., 2007; Sexton et al., 2012). However, until recently model discrepancy has been neglected in the analysis of multi-model ensembles. Rougier et al. (2013) proposed the following framework incorporating a discrepancy between an ensemble of models and the actual climate

$$Z = Y + W \tag{2.14a}$$

$$Y = \mathrm{M}(X) + U \tag{2.14b}$$

$$X_m = \mathrm{M}(X) + R_m \quad \forall\ m = 1, \ldots, M \tag{2.14c}$$

The framework is specified in a general form applicable to the historical climate, future climate, or climate response. As before, W represents the departures of the observations from the actual climate. The $\mathrm{M}(X)$ component represents the model consensus, effectively the ensemble mean. The climates simulated by the models X_m are represented as the consensus $\mathrm{M}(X)$, plus independent and identically distributed mean zero model departures R_m. So all models are treated equally and no model weighting takes place. The actual climate Y is represented by the model consensus $\mathrm{M}(X)$, plus an independent mean zero "ensemble discrepancy" U. The specification is completed by stipulating that the discrepancy U is independent of the model departures R_m.

Rougier et al. (2013) suggest interpreting $\mathrm{M}(X)$ as a representative model, i.e., representative of the climate models, not of the actual climate. With that interpretation in mind, the "ensemble discrepancy" U has a clear interpretation. It represents the effects of shared differences between the models and the actual climate, e.g., the effects of insufficient resolution, and missing processes etc. The associated variance $\mathrm{var}(U)$ quantifies how well the model consensus represents the actual climate. This framework clearly satisfies two of our simple credibility criteria, since the model biases are expected to be correlated

$$\begin{aligned} \mathrm{cov}\left(X_i - Y\ ,\ X_j - Y\right) &= \mathrm{var}\left(U\right) \\ \mathrm{E}\left(\left(\bar{X} - Y\right)^2\right) &= \mathrm{var}\left(U\right) + M^{-1}\,\mathrm{var}\left(R_m\right) \end{aligned}$$

and the mean squared error of the multi-model mean is limited by our the uncertainty due to the discrepancy $\mathrm{var}(U)$. Rougier et al. (2013) note that the "exchangeable" approach is a special case of Equation 2.14 where $\mathrm{var}(U) = \mathrm{var}(R_m)$. They argue that, due to common deficiencies in the models, we would expect the models to be "more like the ensemble mean than the system is like the ensemble mean", i.e., $\mathrm{var}(U) \geq \mathrm{var}(R_m)$. Therefore, our third credibility criterion is also satisfied since

$$\mathrm{var}\left(Y\right) = \mathrm{var}\left(\mathrm{M}\left(X\right)\right) + \mathrm{var}\left(U\right)$$

A similar argument has been used to justify the spread of a multi-model ensemble $\mathrm{var}(R_m)$ as a lower bound on the uncertainty associated with the discrepancy between a perturbed physics ensemble and the actual climate (Sexton et al., 2012).

A subtly different framework was proposed by Chandler (2013)

$$Z = Y + W \tag{2.15a}$$

$$X_m = Y + B + R_m + \varepsilon_m \quad \forall\ m = 1, \ldots, M \tag{2.15b}$$

the common components all have the same interpretations as in the framework of Rougier et al. (2013) above. Once again, the framework is specified in a general form applicable to the historical climate, future climate or climate response. The two frameworks are more easily compared by rewriting Equation 2.15 in terms of the model consensus or representative model (Rougier et al., 2013). Let $\mathrm{M}(X) = Y + B$, then

$$Z = Y + W \tag{2.16a}$$

$$\mathrm{M}\left(X\right) = Y + B \tag{2.16b}$$

$$X_m = \mathrm{M}\left(X\right) + R_m + \varepsilon_m \quad \forall\ m = 1, \ldots, M \tag{2.16c}$$

The B term is described as a "shared discrepancy". It is assumed to be a mean

zero random quantity with associated variance $\mathrm{var}\,(B)$, similar to the "ensemble discrepancy" U of Rougier et al. (2013). The model departures R_m are assumed to be independent and identically distributed with mean zero by default[1]. The ε_m term represents departures due to internal variability and is assumed to have mean zero with model specific variance.

The original formulation of Chandler (2013) in Equation 2.15 is identical to that of Tebaldi and Sansó (2009) in Section 2.5.3. The key difference is that Chandler (2013) interprets B as a random discrepancy rather than a fixed bias. This representation satisfies the first two of our simple credibility criteria since

$$\mathrm{cov}\,(X_i - Y\ ,\ X_j - Y) = \mathrm{var}\,(B)$$

$$\mathrm{E}\left(\left(\bar{X} - Y\right)^2\right) = \mathrm{var}\,(B) + \frac{1}{M}\,\mathrm{var}\,(R_m) + \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{var}\,(\varepsilon_m)$$

so the mean-squared-error of the multi-model mean is limited by the variance of the discrepancy $\mathrm{var}\,(B)$. The precision with which we can know the actual climate Y is also limited by $\mathrm{var}\,(B)$ (Equation 2.16b). Our uncertainty about the actual climate Y (given only information from the models) will span the range of climates simulated by the models provided that $\mathrm{var}\,(B) \geq \mathrm{var}\,(R_m)$, as argued above. Therefore, our third criteria and the intuition of Tebaldi and Sansó (2009) are both satisfied. Rougier et al. (2013) note that the generalised "truth plus error" framework proposed by Tebaldi and Sansó (2009) can be considered a special case of the framework proposed by Chandler (2013).

2.5.6. Ensemble regression

In Section 2.4, we introduced the concept of an "emergent constraint" - a physical constraint in the climate system that manifests as a correlation between the projected responses and the historical climates simulated by an ensemble of climate models. Most studies of emergent constraints have assumed a linear relationship between the future climate or the climate response and the historical climate (e.g., Hall and Qu, 2006; Boé et al., 2009; Cox et al., 2013). This approach was formalised by Bracegirdle and Stephenson (2012) under the name "ensemble regression", so that

$$x_{Rm} = \beta + \lambda\,(x_{Hm} - \bar{x}_H) + R_m \tag{2.17}$$

where x_{Rm} and x_{Hm} represent the climate response and historical climate simulated by model m, and $\overline{x}_H$ is the multi-model mean historical climate of the models. The

[1] Chandler (2013) notes that $\mathrm{var}\,(R_m)$ quantifies "the propensity for the [...] simulator to deviate from the simulator consensus" and that "the use of simulator specific covariance matrices [...] provides some flexibility to accommodate outlying simulators".

intercept parameter β represents the expected response of the ensemble, and the slope parameter λ characterises the emergent constraint. The departures R_m are assumed to be independent and identically distributed mean zero normal random variables. Bracegirdle and Stephenson (2012) estimate x_{Rm} and x_{Hm} by the mean of all available initial condition runs from model m. This reduces the impact of initial condition uncertainty. Therefore, the variance $\mathrm{var}\,(R_m)$ is primarily a measure of model uncertainty.

The actual climate response y_R is estimated by substituting observations z for the modelled historical climate x_{Hm} on the right hand side of Equation 2.17, as illustrated in Figure 2.1. Bracegirdle and Stephenson (2012) estimate the uncertainty about y_r as the uncertainty associated with the response of a new model, given knowledge of its historical climate

$$\mathrm{var}\,(y_R) = \mathrm{var}\left(\hat{\beta} + \hat{\lambda}\,(z - \bar{x}_H) + R_m\right) = \mathrm{var}\,(R_m)\left(1 + \frac{1}{M} + \frac{(z - \bar{x}_H)^2}{\sum_{m=1}^{M}(x_{Hm} - \bar{x}_H)^2}\right) \tag{2.18}$$

where $\hat{\beta}$ and $\hat{\lambda}$ are the maximum likelihood estimates of the parameters. As the ensemble size increases, the uncertainty about the actual climate response y_R is limited by $\mathrm{var}\,(R_m)$. Therefore, the third of our simple credibility criteria is satisfied. Note that $\mathrm{var}\,(R_m)$ quantifies the conditional uncertainty about the response of a model, given its historical climate. This is smaller than the full spread of the model responses, as illustrated by the prediction internal in Figure 2.1. So our uncertainty about the actual climate response is reduced by ensemble regression. Since no probabilistic description is specified for the historical climates of the models x_{Hm}, we cannot say anything about our other two criteria. The basic methodology proposed by Bracegirdle and Stephenson (2012) has since been extended to the relationship between the climate response of one variable, and the historical state of several predictor variables using multiple linear regression Karpechko et al. (2013).

A similar methodology to ensemble regression was proposed by Räisänen et al. (2010), except that instead of regressing directly on the model climates and climate responses, they regress on the differences between all possible combinations of model climates and responses

$$x_{Ri} - x_{Rj} = \beta' + \lambda'\,(x_{Hi} - x_{Hj}) + \epsilon_{ij} \tag{2.19}$$

Bracegirdle and Stephenson (2012) point out that the two methods are closely related. The difference between the climate responses of two models in ensemble regression (Equation 2.17) is

$$x_{Ri} - x_{Rj} = \lambda\,(x_{Hi} - x_{Hj}) + R_i - R_j$$

which is equivalent to Equation 2.19 with $\beta' = 0$. However, the formulation in terms of model differences cannot directly be used for projection in the same way that ensemble regression can. Instead, Räisänen et al. (2010) estimate the actual climate response using a heuristic average of the model responses x_{Rm}, weighted by the historical biases of the models compared to the observations. This complicates the interpretation of the projection and its associated uncertainty, which must be estimated by cross validation.

An alternative method analysing emergent relationships has been proposed using maximum covariance analysis to identify spatial patterns of correlation between the models' responses and their historical states (Abe et al., 2011). Projections based on observations are possible, but additional assumptions are required to obtain estimates of the associated uncertainties.

2.5.7. Constant relationship methods

The family of methods described in this section originated in seasonal climate forecasting and have subsequently been adapted for the estimation of long term climate change. The underlying methodology was first proposed by Krishnamurti et al. (1999, 2000) who suggested treating the model outputs x_m as predictors for the actual climate y in order to obtain a set of "optimal" weights for the individual models by multiple linear regression

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.20}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$ is a vector of N observations, and $\boldsymbol{\beta} = (\beta_1, \ldots \beta_M)^T$ is a vector of M regression coefficients. The design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$ is composed of vectors of retrospective forecasts of the observations $\mathbf{y}$, where $\mathbf{x}_m$ is the vector of forecasts from model m. The departures $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$ arise due to a combination of observation error, internal variability and model error. Standardising the forecasts $\mathbf{x}_m$ by removing the mean bias from each model will reduce the contribution due to model error. The intercept term β_0 can be neglected if the forecasts have been standardised (DelSole, 2007).

The assumption when used for projection is that each model maintains a constant relationship with the actual climate through the regression coefficients. However, Kharin and Zwiers (2002) found that multi-model regression tends to be overconfident when applied to new data, due to the relatively small number of observations that are usually available to estimate the regression coefficients. Yun et al. (2003) addressed this problem by performing the regression on only the first dimension suggested by a principle component analysis of the model forecasts. Peña and van den

Dool (2008) used the related technique of ridge-regression (Hoerl and Kennard, 1970). Ridge-regression and the various simpler regression estimates proposed by Kharin and Zwiers (2002) can all be viewed from a Bayesian perspective as restrictions on the prior distributions of the regression coefficients $\boldsymbol{\beta}$ (DelSole, 2007).

All of the developments described above were based on applications to seasonal climate forecasting. However, multi-model regression methods have also been applied to long term climate change. The key difference is that when fitting the regression coefficients $\boldsymbol{\beta}$, the design matrix $\mathbf{X}$ is made up of simulations of the historical climate, rather than retrospective forecasts. The number of observations available for fitting is still small, however. Bishop and Abramowitz (2013) used constrained regression, similar to the ridge regression variants described by Peña and van den Dool (2008). While Greene et al. (2006) took a Bayesian approach similar to DelSole (2007).

Our simple credibility criteria can be evaluated by reference to properties of constrained multiple linear regression. Greene et al. (2006) assume that the regression coefficients $\boldsymbol{\beta}$ arise from a multivariate normal distribution. The off-diagonal elements of the associated covariance matrix represent correlations between the models, so our first criteria is satisfied. In unconstrained linear regression the sum of squared errors compared to the observations used for fitting $\mathbf{y}_H$ will always decrease as the number of predictors (models) increases. However, the multi-model mean is equivalent to constrained regression where $\beta_m \approx 1 \; \forall \; m$. We have already seen that the mean squared error of the multi-model mean will only converge to zero if there is no shared bias or discrepancy between the models and the actual climate. So our second criteria is partially satisfied. The predictive uncertainty for the future climate $\mathbf{y}_F$ (and hence the climate response) is limited by $\mathrm{var}\,(\varepsilon)$, similar to ensemble regression in Equation 2.18. However, $\mathrm{var}\,(\varepsilon)$ itself has the sample variance of the observations used for fitting $\mathbf{y}_H$ as its upper limit. So our third criteria is not guaranteed since there is no reason that we should expect $\mathrm{var}\,(\mathbf{y}_H) \geq \mathrm{var}\,(\mathbf{x}_{Rm})$.

Similar approaches are also used to make projections of future climate based on detection and attribution methods (Allen et al., 2000; Stott and Kettleborough, 2002; Stott et al., 2006a,b; Stott and Forest, 2007). In those studies, the design matrix $\mathbf{X}$ in Equation 2.20 is composed of response patterns ("fingerprints") to different combinations of radiative forcing (greenhouse gases etc.) simulated by a single climate model. The regression coefficients $\boldsymbol{\beta}$ represent the contribution of each pattern of forcing to the observed climate change. Projection by this method assumes that the contribution of each pattern of forcing remains constant in the future, i.e., maintains a constant relationship with the actual climate response (Kettleborough et al., 2007). Fingerprints computed from a future forcing scenario are then substituted into $\mathbf{X}$ to obtain projections $\mathbf{y}_F$.

2.5.8. Mixed methods

The expected value of the actual climate response Y_R in the "truth plus error" approach can be written as a weighted combination of the model responses (Tebaldi et al., 2005). In this respect, the "truth plus error" and "constant relationship" approaches are not dissimilar. Buser et al. (2009) proposed a framework that combined elements of both approaches

$$Z_H = Y_H + W_H \tag{2.21a}$$

$$Z_R = Y_R + W_R \tag{2.21b}$$

$$X_{Hm} = Y_H + R_{Hm} + \varepsilon_{Hm} \tag{2.21c}$$

$$X_{Rm} = Y_R + R_{Rm} + \varepsilon_{Rm} \tag{2.21d}$$

where Z_H and Z_R are the observed climate and yet to be observed climate response, the expected values of which are Y_H and Y_R. The departures W_H and W_R are assumed to be due to natural variability. The model climates X_{Hm} and climate responses X_{Rm} are modelled as the actual climate Y, plus a departure due to model error R_m, plus a departure due to internal variability ε_m. This is clearly a "truth plus error" framework and strongly resembles the methods proposed by Tebaldi et al. (2005) and Tebaldi and Sansó (2009). However, Tebaldi et al. (2005) assumed that the response of model m was correlated with its historical climate X_{Hm}, i.e., an emergent relationship. Instead, Buser et al. (2009) assume that the response of model m is proportional to the response of the actual climate Y_R

$$\mathrm{E}\left(R_{Rm}\right) = \left(\phi_m - 1\right) Y_R \quad \text{where} \quad \mathrm{var}\left(\varepsilon_{Hm}\right) = \phi_m^2 \, \mathrm{var}\left(W_H\right)$$

so model m will tend to over (under) estimate the climate response by an amount that is proportional to its over (under) estimation of the internal variability, i.e., if the temperature of warm (cool) years in model m tend to be too warm (cool) by a factor of ϕ_m and the climate warms (cools), then model m will simulate a mean response that is also too warm (cool) by a factor of ϕ_m. So this could also be classed as a constant relationship framework. Never the less, it fails to satisfy any of our simple credibility criteria for the same reasons as all the other "truth plus error" frameworks described in Section 2.5.3.

2.5.9. Discussion

Other methods of combining projections from multiple climate models are possible (e.g., Min and Hense (2006) used Bayesian Model Averaging (Hoeting et al., 1999)), but the approaches summarised in the preceding sections are the most common.

Although widely used at shorter time scales, constant relationship methods are rarely applied to century scale climate projection. Stott and Forest (2007) found that the assumption of constant relative contributions by different forcing components inherent in was unlikely to hold over long time scales. A similar argument can also be made regarding the multi-model regression frameworks proposed by Greene et al. (2006) and Bishop and Abramowitz (2013). As the balance between different processes alters in a changed climate, it is unlikely that models which may represent those processes very differently will maintain a constant relationship with the actual climate. Buser et al. (2009) compared the constant relationship assumption with a constant bias assumption similar to Tebaldi and Sansó (2009) and concluded that "the "constant bias" assumption is the more natural assumption in longer-term climate change studies".

The frameworks based on reliability ensemble averaging (Giorgi and Mearns, 2002) have been criticised for their use of a "convergence" criteria (Lopez et al., 2006). Given the possibility of shared errors due to processes missing from or poorly represented by all models, it may be unwise to attach additional weight to models simply because they agree with each other. However, Tebaldi and Knutti (2007) point out that it is implicit in the common practice of excluding models whose responses are extreme compared to the rest of the ensemble.

The "exchangeable" approach is attractive for its simplicity. However, it is likely to be an oversimplification of the true situation. In Section 2.2, it was argued that the models that make up a multi-model ensemble cannot be considered independent, and are really a sample of "best guesses" calibrated to the recent climate. As a result, they are unlikely to sample the full range of structural uncertainty about the future climate response. Therefore, an "exchangeable" approach is likely to underestimate that uncertainty. However, Sanderson and Knutti (2012) argue that unless the response is constrained by the observations, then even an ensemble of "best guesses" will diverge and sample a much larger range of uncertainty in the future.

Provided that the underlying physical processes are understood, emergent relationships provide important opportunities to constrain projections of future climate change using observations of recent climate. However, by estimating the uncertainty associated with the actual climate response by the uncertainty about the response of a new model, Bracegirdle and Stephenson (2012) implicitly assume that the actual climate will differ from the models in the same way that the models differ from each other, i.e., they are sampled from the same population or "exchangeable". We have argued above that this assumption may be too strong.

The frameworks proposed by Chandler (2013) and Rougier et al. (2013) generalise

the "truth plus error" and "exchangeable" approaches, respectively. Both include a discrepancy term that explicitly captures the idea that the climate models are fundamentally different from the actual climate, and both satisfy all of our simple credibility criteria. Of the existing methods for synthesising projections from multi-model ensembles, these two seem the most promising. Neither incorporates all the features we might wish to see, however. Rougier et al. (2013) do not explicitly account for internal variability in the climates simulated by the models, and neither framework incorporates the ability to constrain future projections using emergent relationships. Further, none of the frameworks discussed in this section attempts to separate the effects of sampling uncertainty and measurement error.

2.6. The CMIP5 multi-model ensemble

The methodology developed in subsequent chapters will be illustrated by application to climate variables simulated by the ensemble of models participating in the fifth phase of the Coupled Model Intercomparison Project, CMIP5 (Taylor et al., 2012). The ensemble includes simulations from more than 40 models, submitted by more than 20 modelling centres around the world. The models included in the comparison are all atmosphere-ocean general circulation models (AOGCMs). General circulation models represent the equations of motion on a sphere. Coupled atmosphere-ocean models include both atmosphere and ocean components, and the interactions between them. Most modern AOGCMs also include interactive land surface and sea-ice components. CMIP5 is the first CMIP in which some of the models also include interactive carbon cycle components. These models are known as Earth System Models (ESMs).

The experiments included in CMIP5 are designed to address key knowledge gaps, and to provide "a framework for coordinated climate change experimentation [...] over the next several years" (Taylor et al., 2012). The outputs from CMIP5 form the basis for most of the assessment contained in the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) (Stocker et al., 2013). A mixture of long-term (century time scale), and near-term (10-30 year) climate change experiments are included. The focus of this thesis is on long-term climate change. Four future scenarios are included in the suite of experiments defined for CMIP5. These scenarios are known as "representative concentration pathways" (RCPs) (Moss et al., 2010). They are identified by the expected increase in radiative forcing (Wm^{-2}) at the year 2100, according to the integrated assessment model that generated the scenario, e.g., RCP4.5 is expected to produce an increase in radiative forcing of $4.5\mathrm{Wm}^{-2}$ by 2100. The four chosen scenarios were selected to span the range of scenarios of radiative forcing and greenhouse gas emissions

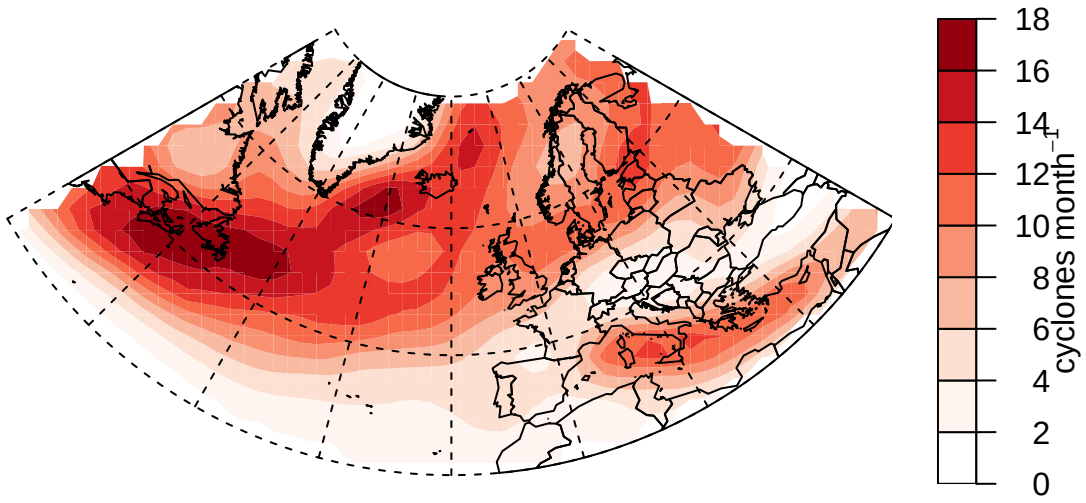

Figure 2.3.: Winter (December-January-February) extra-tropical cyclone track density in the North Atlantic storm track computed from ERA-Interim reanalysis data using Hodges' TRACK algorithm. The track density is the mean number of cyclones passing within 5° of a grid point each month.

considered plausible at the time (Moss et al., 2010).

A separate set of experiments were defined for the Earth System Models based on the same scenarios, but forced by observed or prescribed emissions rather than concentrations of greenhouse gases. Interactive representations of the carbon cycle enable Earth System Models to compute the concentrations based on the emissions. This enables additional assessment of poorly understood feedback mechanisms in the carbon cycle. Outputs from these additional experiments are not included in any of the analysis included in this thesis.

Due to the associated computational expense, modelling centres are not required to perform every experiment included in the CMIP5 design. A set of core experiments were defined, including at least one initial condition run each of a historical scenario, the RCP4.5 midrange mitigation scenario, and the RCP8.5 high emissions scenario (Taylor et al., 2012). The historical scenario is forced by observed concentrations of greenhouse gases over the twentieth century, but the atmosphere and ocean etc. are not initialised from observed states. Therefore, the historical scenario should be expected to reproduce the climate of the twentieth century (the distribution of weather), but not individual events (e.g., a particular warm summer) due to the internal variability in the models and natural variability in the Earth system.

2.7. Extra-tropical cyclone frequency in the North Atlantic

The main data set that will be analysed in the chapters that follow, concerns the frequency of extra-tropical cyclones over the North Atlantic, Europe, and the Mediterranean (Figure 2.3). Extra-tropical cyclones are primarily of interest due to the potential for large societal impacts. The precipitation associated with cyclones has an important role in maintaining the water supply of Europe and the Mediterranean. If the amount of precipitation supplied by extra-tropical cyclones were to be reduced due to climate change, then this could have serious consequences for water supplies in vulnerable regions. Cyclones are also associated with a range of hazards, including high winds and flooding. If cyclone activity increases, then the storm damage and widespread flooding that affected large parts of the UK between December 2013 and February 2014 (Slingo et al., 2014) may become more frequent. The other reason for studying extra-tropical cyclones is that CMIP5 (Taylor et al., 2012) is the first large, coordinated multi-model ensemble to specifically include high frequency output so that individual cyclones can be identified and tracked through their life cycles.

Extra-tropical cyclones are mobile low pressure systems occurring in the mid-latitudes. They tend to organise themselves into loosely defined *storm tracks* spanning the North Atlantic, North Pacific, and Southern oceans. More localised storm tracks also exist in the Mediterranean and over Mongolia and Northern China. Storm tracks exist on the boundaries between the warm tropical air masses, and the cold polar air masses. It is from the strong temperature gradients that exist in the mid-latitudes that extra-tropical cyclones draw their energy (Charney, 1947; Eady, 1949). The storm tracks are most active in winter when the meridional temperature contrast between the tropics and the poles is strongest. On the east coast of the United States, the temperature gradients are further enhanced by the contrast between the cold land and the warm waters of the Gulf stream. This leads to an intense region of cyclogenesis over the ocean between Cape Hatteras and Cape Cod (Hoskins and Hodges, 2002). The North Atlantic storm track is sometimes described as having a main branch towards Iceland and Norway, and a secondary zonal branch towards Central Europe (Blender et al., 1997). The UK sits between the two branches. However, it is more strongly effected by the main branch since the strongest winds and precipitation tend to occur to the south-east of the center of the cyclone (Bengtsson et al., 2009).

It is this dependence on temperature gradients that make it difficult to predict how cyclone activity may be affected by climate change. In the lower troposphere, the equator-to-pole temperature gradient is projected to decrease, due to enhanced

warming in the Arctic (Collins et al., 2013). This may lead to a corresponding decrease in the frequency or intensity of extra-tropical cyclones. On the other hand, the temperature gradient in the upper troposphere is likely to be enhanced by climate change, due to the maxima in warming in the tropics (Collins et al., 2013, Figure 12.12). This in turn may increase the energy available to extra-tropical cyclones. A third factor is the energy derived from the latent heat released by precipitation within the cyclones. A warmer climate means more water vapour available for precipitation (Boucher et al., 2013). This may also act to increase the frequency or intensity of extra-tropical cyclones. The balance between these competing factors is difficult to predict from theoretical arguments. Therefore, projections of cyclone activity derived from general circulation models are necessary to assess future changes.

As noted above, CMIP5 is the first large multi-model ensemble to include the high frequency (6 hourly) output required to identify individual cyclones. Previously, it was common to compute the variance of the 2-6 day bandpass filtered daily mean sea level pressure for each grid box. Filtering the data removes low frequency variations due to seasonal forcing, as well as residual high frequency variations due to the diurnal cycle (Blackmon, 1976). The remaining signal should represent variations caused by transient weather systems, i.e., cyclones. However, this simple method will capture variations due to both high and low pressure systems, as well as features that are not of interest such as heat lows. If high frequency data are available, then individual cyclones can be identified, and tracked through their life cycles. A variety of objective (automated) tracking methods have been proposed. Most are based on identifying minima in mean sea level pressure (e.g., Murray and Simmonds, 1991) or maxima in the relative vorticity (a measure of the spin of a fluid) either at sea level or some higher level in the troposphere (e.g., Hodges, 1999). This introduces an additional source of uncertainty, since different tracking methods may identify different features. Recent comparisons based on historical data have shown that the majority of tracking methods are consistent in their ability to identify and track strong cyclones (Neu et al., 2013). They are less consistent for weak systems, due to differences in the criteria used to define a cyclone (Neu et al., 2013). Additional comparisons using climate model output have shown that the climate change signal is also robust to the tracking method used (Ulbrich et al., 2013).

As discussed earlier, good performance in reproducing historical climate is a necessary (but not sufficient) requirement for confidence in future projections (Knutti et al., 2010b; Stephenson et al., 2012). The authors of the IPCC fifth assessment report (AR5) concluded that "models are able to capture the general characteristics of storm tracks and extra-tropical cyclones" (Flato et al., 2013). High resolution climate models have even been shown to capture the internal structure cyclones (Bengtsson et al., 2009; Catto et al., 2010). However, one problem in particular has

persisted in climate models for several generations. König et al. (1993) found the North Atlantic winter storm track in the ECHAM2 model to be too active in the zonal branch and not active enough in the main branch. This finding was repeated in 11 out of 13 models participating in the AMIP1 comparison with too many cyclones passing over the UK and Central Europe (Lambert et al., 2002). This problem persists in some of the CMIP5 models (Zappa et al., 2013a; Chang et al., 2012). Comparing the variance statistics for the mean sea level pressure shows that the CMIP5 models are generally less biased in the North Atlantic storm track compared to the previous generation of models included in CMIP3 (Zappa et al., 2013a). The CMIP5 models also tend to underestimate the intensity of extra-tropical cyclones in the North Atlantic during the winter season (Chang et al., 2012; Colle et al., 2013). There is a general tendency to simulate too many weak cyclones *and* not enough strong cyclones (Zappa et al., 2013a). So while model performance in simulating extra-tropical cyclones is improving, projections should still be interpreted with caution.

Studies of climate change in the storm tracks consistently project a poleward shift of several degrees in the Southern hemisphere by the end of the next century (Bengtsson et al., 2006; Gastineau and Soden, 2009; Chang et al., 2012). Applying a simple tracking algorithm to the daily data from the CMIP3 models suggested an overall decrease in the number of extra-tropical cyclones in the Northern Hemisphere (Lambert and Fyfe, 2006). Studies of individual models showed decreases in cyclone frequency in the Mediterranean and on the southern flank of the North Atlantic storm track, accompanied by increases over the UK and Ireland (Geng and Sugi, 2003; Bengtsson et al., 2006). This led the authors of the IPCC fourth assessment report (AR4) to conclude that models project "a poleward shift of storm tracks in both hemispheres by several degrees of latitude" (Meehl and Coauthors, 2007). However, further analysis of the CMIP3 models and the newer CMIP5 models led to the conclusion that "it is unlikely that the response of the North Atlantic storm track is a simple poleward shift" in the IPCC fifth assessment report (Christensen et al., 2013). Instead, Zappa et al. (2013b) describe the pattern of projected change as "tri-polar", with decreasing cyclone frequency on both the northern and southern flanks of the North Atlantic storm track and in the Mediterranean, and increasing activity over UK, Ireland, Denmark and southern Norway. The decrease on the northern flank of the North Atlantic storm track was also evident in the earlier single model studies (Geng and Sugi, 2003; Bengtsson et al., 2006). However, it seems to have been ignored due to the simpler poleward shifts projected in the Southern Hemisphere and North Pacific storm tracks. Similar tri-polar patterns have been noted in other single model studies, so the projection seems to be robust (Pinto et al., 2007; McDonald, 2011; Ulbrich et al., 2013).

Most studies also predict a similar pattern of increasing frequency when only strong cyclones are considered, except in the Mediterranean where no decrease in the number of strong cyclones is projected (Geng and Sugi, 2003; McDonald, 2011; Mizuta, 2012; Zappa et al., 2013b). Only Bengtsson et al. (2006) found evidence of a decrease in the number of intense cyclones in the Mediterranean. Elsewhere, the magnitude of the changes are smaller than for all cyclones, suggesting that changes in cyclone frequency will be dominated by weak systems that probably wouldn't be classified as storms. The term cyclone is used to describe any cyclonic circulation system, where as storm is sometimes reserved for systems leading to strong winds or heavy rain at the surface. There is no widely agreed definition for either a cyclone or a storm, and this can lead to differences between tracking methods and between studies, particularly in regions where cyclones usually begin or end their life cycles.

Several studies have suggested the potential for emergent constraints on extratropical cyclone activity. The climate change response of the North Atlantic storm track has been linked to changes in the meridional overturning circulation in the ocean Woollings et al. (2012). Climate models that simulate strong overturning in the historical scenario tend to simulate a larger reduction in overturning in the future Gregory et al. (2005). The emergent relationship on the overturning circulation could be used to constrain the response of the storm track. It has also been shown that the climate response in the storm track is indeed correlated with the change in equator-to-pole temperature gradients at both low and high levels of the atmosphere, as argued above (Harvey et al., 2013). A number of emergent constraints have been found on polar near surface temperature (e.g., Hall and Qu, 2006; Boé et al., 2009; Bracegirdle and Stephenson, 2012). So, it may be possible to constrain the climate response of the storm tracks using the relationship between historical and future polar temperatures. Chang et al. (2012, 2013) actually found evidence in both the CMIP3 and CMIP5 ensembles that climate models with more active storm tracks in the historical scenario tend to simulate stronger decreases in activity in future scenarios in the Northern hemisphere. Therefore, it may be possible to constrain projections of cyclone frequency directly from the historical cyclone activity simulated by the models.

2.8. Summary

This chapter has reviewed the issues associated with interpreting the outputs of ensembles of multiple climate models, and the various methodologies that have been proposed for synthesising projections. Of the statistical frameworks proposed to date, only two satisfy all three of the simple credibility criteria identified from the literature. Many confound differences between climate models with departures due

to internal variability. Some fail to account for uncertainty in observations of recent climate when combining model output with observational data. Only a few methods explicitly account for possible correlations between the climate responses simulated the models and their historical climates. In this thesis, we construct and derive checking procedures for a probabilistic description of the outputs of an ensemble of climate models that includes all these features. In the first instance we focus on quantifying uncertainty due to internal variability. The framework is then extended to quantify structural uncertainty due to differences between models, and later to include the estimation of emergent relationships. In the final chapter, the ensemble is linked to the actual climate via an ensemble discrepancy similar to that suggested by Rougier et al. (2013) and Chandler (2013). The developments in each chapter are illustrated by application to the estimation of the climate change response of cyclone frequency in the North Atlantic, using an extended version of the data analysed by Zappa et al. (2013b) and Sansom et al. (2013).

3. Analysis of variance methods

3.1. Introduction

The goal of this thesis is to make credible inferences about the future climate of the Earth system based on the outputs of an ensemble of climate models. The emphasis in the early chapters is on formulating and checking a probabilistic description of the outputs of an ensemble of climate models. The relationship between the climate models and the Earth system is considered in the context of the paradigms discussed in Chapter 2, but not defined explicitly. The relationship between the models and the Earth system is considered in detail in Chapter 6.

Yip et al. (2011) showed how a simple ANOVA framework could be used to quantify the relative contributions of model uncertainty, scenario uncertainty, and internal variability in climate projections. However, ANOVA frameworks can be used to estimate parameters in statistical frameworks as well as relative uncertainties. Simple ANOVA frameworks have been used to analyse projections from ensembles of regional as well as global climate models (Ferro, 2004; Hingray et al., 2007). Other studies have used the basic ANOVA methodology as the basis for more complex frameworks (Sain et al., 2011; Kang and Cressie, 2013).

In this chapter, a hierarchy of ANOVA frameworks is introduced that can be used to analyse multi-model ensembles of climate simulations. The ANOVA frameworks make explicit one simple set of assumptions that lead naturally to the usual "one model, one vote" estimate of the actual climate response. Standard theory permits the construction of confidence intervals for the true value of the expected climate response of the ensemble. However, it is argued that reporting such intervals based on the "one model, one vote" estimate as representative of the actual climate response would be misleading. If the contribution from model uncertainty is small, i.e., the models agree on the climate response, then an alternative weighted average is derived for which interval estimates may be reported with confidence. Hypothesis tests are derived for evidence of model agreement and non-zero climate response. Using these tests, it is shown that the CMIP5 models are actually in good agreement on the climate response in the North Atlantic storm track. Power analysis of the test for non-zero climate response is used to address the question of whether the

CMIP5 ensemble is large enough to reliably detect future climate change.

3.2. The multi-model mean

Any long term climate projection is conditional on the forcing scenario that is used to generate it. The focus of this thesis is on combining information from multiple models. Therefore, all analysis will be restricted to ensembles containing only one historical (H) and one future (F) scenario, i.e., $s \in \{H, F\}$ where s denotes the scenario. Let X_{smr} be a random variable representing a climate statistic (e.g., a 30-year average) from initial condition run r of scenario s simulated by climate model m. Some studies estimate the expected climate response of a model m using the difference between only one run from each scenario (e.g., Tebaldi et al., 2005; Collins et al., 2013). The output of individual runs will differ from the expected climate of the model due to unforced internal variability. Alternatively, the expected response of model m (e.g., Bracegirdle and Stephenson, 2012; Meehl and Coauthors, 2007, Figure 10.5) may be estimated by

$$\overline{x}_{Fm.} - \overline{x}_{Hm.}$$

where $\overline{x}_{sm.}$ is the mean over all runs of scenario s from model m

$$\overline{x}_{sm.} = \frac{1}{N_{sm}} \sum_{r=1}^{N_{sm}} x_{smr}$$

and N_{sm} is the number of runs of scenario s from model m. The weak law of large numbers guarantees that $\overline{x}_{sm.}$ will converge to the expected climate of model m in scenario s, regardless of the distribution of X_{smr}.

Whether we adopt a "truth plus error" or an "exchangeable" view of the relationship between the climate models and the Earth system, the mean climate response of an ensemble of climate models is a sensible estimate of the actual climate response. In the "truth plus error" approach, the expected response of the ensemble is assumed to be equal to the actual climate response. In the "exchangeable" paradigm, the expected response of the ensemble is the expectation of the distribution from which both the model responses and the actual response are drawn. The sample mean of the mean responses simulated by the models is a natural estimate of the expected response of the ensemble. A general multi-model mean estimate of the expected climate response is given by

$$\frac{1}{w_{F.}} \sum_{m=1}^{M} w_{Fm} \overline{x}_{Fm.} - \frac{1}{w_{H.}} \sum_{m=1}^{M} w_{Hm} \overline{x}_{Hm.} \tag{3.1}$$

where M is the number of models in the ensemble and w_{sm} is a weight applied to the mean of model m in scenario s, and

$$w_{.s} = \sum_{m=1}^{M} w_{sm}$$

is the sum of the model weights in scenario s. The most commonly used estimate of the future climate response is the equally weighted multi-model mean or "one model, one vote" estimate

$$\frac{1}{M} \sum_{m=1}^{M} (\overline{x}_{Fm.} - \overline{x}_{Hm.}) \tag{3.2}$$

equivalent to

$$w_{Hm} = w_{Fm} = 1 \quad \forall\ m = 1, 2, \ldots, M \tag{3.3}$$

By assigning each model equal weight, this estimate treats each model as an equally valid representation of the climate system.

Although simple and easy to compute, ad hoc multi-model means have a number of problems. Unless the model weights reflect the true predictive performance of the models, then the projections are likely to be less accurate than if the equal weight were given to all models (Weigel et al., 2010). Therefore, we will not attempt to derive performance based model weights. Even when the equally weighted multi-model mean is reported, the underlying assumptions behind the estimate are often not made clear, and therefore cannot be checked. No measure of the uncertainty associated with the estimate is usually provided, and so confidence intervals cannot be constructed for the climate response. Further, as an arithmetic mean, the estimate may be strongly influenced by models or runs that are outlying compared to the rest of the ensemble.

3.3. ANOVA frameworks

In Sansom et al. (2013), a family of ANOVA frameworks was introduced for the analysis of multi-model ensemble climate change experiments. This section describes the structure of those frameworks, and the interpretation of the various parameters.

3.3.1. A two-way ANOVA framework with interactions

In Appendix A.1, the "one model, one vote" estimate (Equation 3.2) of the climate response is shown to be equivalent to the maximum likelihood estimate $\hat{\beta}_F$ of the expected climate response of an ensemble of climate models from the following

ANOVA framework

$$\begin{aligned} x_{smr} &= \mu + \alpha_m + \beta_s + \gamma_{sm} + \varepsilon_{smr} \\ \varepsilon_{smr} &\overset{iid}{\sim} N\left(0, \sigma^2\right) \end{aligned} \tag{3.4}$$

subject to the constraints that $\sum_{m=1}^{M} \alpha_m = 0$, $\beta_H = 0$, $\gamma_{Hm} = 0 \; \forall \; m = 1, \ldots, M$ and $\sum_{m=1}^{M} \gamma_{Fm} = 0$. The main effects μ and β_F represent the expected historical climate and climate response of the ensemble respectively. The α_m terms represent the departure of the expected historical climate of model m from the expected historical climate of the ensemble (μ). The interaction terms γ_{Fm} represent the departure of the expected climate response of model m from the expected climate response of the ensemble (β_F). The constraints on the α_m and γ_{Fm} terms ensure that the historical climates and climate responses of the the individual models are centred on the expected historical climate and climate response of the ensemble. The random component ε_{smr} represents internal variability and is assumed to be normally distributed. Internal variability in this context is unforced natural variability in the system, sampled by starting each run with slightly different initial conditions.

A total of $P = 2M$ parameters (excluding the variance σ^2) must be estimated in order to fit the framework described by Equation 3.4. One parameter is estimated for the expected historical climate (μ) and one for the expected climate response (β_F). However, there are only $2M$ group means in the data, one for each model-scenario pair. In order to make all the model specific effects identifiable, the α_m and γ_{Fm} terms are constrained to be centred on μ and β_F respectively. Therefore, only $M - 1$ of each must be estimated. This leaves $N_{\cdot\cdot} - 2M$ degrees of freedom to estimate the internal variability (σ^2), where $N_{\cdot\cdot} = \sum_m \sum_s N_{sm}$. In a small ensemble, the remaining degrees of freedom may be small and so the precision of the estimates will be low. If only one run is available from each model for each scenario, then $N_{\cdot\cdot} = 2M$ and it is not possible to estimate the internal variability.

Additional assumptions are required in order to justify the estimated response of the ensemble $\hat{\beta}_F$ as an estimate of the actual climate response. In the "truth plus error" approach, the expected historical climate and future climate response of the models μ and β_F are assumed to coincide with the actual historical climate and climate response. However, there is empirical evidence that shared errors may exist so that the models are not centred on the actual climate (Knutti et al., 2010b). A less restrictive assumption is that the models are centred on some ensemble mean, which may not coincide with the actual climate, i.e., there is a discrepancy between the expected climate of the models and the actual climate (Chandler, 2013; Rougier et al., 2013). If we are only interested in the climate response, then we might be willing to assume that any discrepancy between the expected climate of the models and the actual climate is constant over time, i.e, there is no discrepancy between

the expected response of the models β_F and the actual climate response. This assumption may still be optimistic (Christensen et al., 2008; Buser et al., 2009), however it is more justifiable than the assumption of no discrepancy at all.

The interaction terms γ_{Fm} complicate the interpretation of the ensemble expected climate response (β_F). If the models all simulate different responses, then it is hard to know how the actual climate will respond. The model specific terms α_m and γ_{Fm} represent model differences that give rise to additional uncertainty in the historical climate and climate response. The relative contributions of these sources of uncertainty to the total variability in the ensemble are quantified in Section 3.5. However, only the uncertainty due to internal variability is quantified absolutely, by the parameter σ^2. The structural uncertainty in the climate response has been modelled out by the γ_{Fm} terms. Reporting a confidence interval for the true value of β_F based on the framework of Equation 3.4 would neglect the contribution from structural uncertainty and therefore be overconfident.

3.3.2. A simpler two-way ANOVA framework

If the models all simulate the same climate response (i.e., $\gamma_{Fm} = 0 \; \forall \; m = 1, \ldots, M$), then estimating the γ_{Fm} terms in Equation 3.4 is unnecessary. Estimating a systematic component where none exists introduces additional variance into the framework. In that case, a simpler two-way ANOVA framework should provide more precise estimates

$$\begin{aligned} x_{smr} &= \mu + \alpha_m + \beta_s + \varepsilon_{smr} \\ \varepsilon_{smr} &\overset{iid}{\sim} N\left(0, \sigma^2\right) \end{aligned} \tag{3.5}$$

with the constraints that $\sum_{m=1}^{M} \alpha_m = 0$ and $\beta_H = 0$. The effects have the same interpretation as in Equation 3.4. However, the maximum likelihood estimates are different. In Appendix A.2, it is shown that the maximum likelihood estimate of the ensemble expected climate response ($\hat{\beta}_F$) from the two-way framework is a weighted average of the model mean responses with weights

$$w_{Hm} = w_{Fm} = \frac{N_{Hm} N_{Fm}}{N_{Hm} + N_{Fm}} \tag{3.6}$$

The weights depend on both the number of historical runs and the number of future runs. For a model to obtain a high weighting, it is not sufficient to have a large number of historical runs (i.e., $w_{Hm} = w_{Fm} \approx N_{Fm}$ if $N_{Hm} \gg N_{Fm}$). This contradicts advice not to weight models based on the number of runs they contribute to the ensemble (Knutti et al., 2010a). If the models do all simulate the same climate response then it makes sense that those models that contribute the most runs receive

the most weight.

Without the interaction terms, there are only $P = M+1$ parameters to be estimated. Therefore, the precision of the parameter estimates should increase, since there are $M - 1$ additional degrees of freedom remaining to estimate the internal variability σ^2 compared to the framework with interactions. However, if the models do not all simulate the same climate response, then a systematic component is missing from the framework and the estimates will be biased (Appendix A.4). If the missing effects are large, then the precision of the estimates will decrease dramatically as the estimate of the internal variability (σ^2) will be inflated to compensate. Therefore, the two-way framework should only be used when we are confident that the models do all simulate a similar climate response.

If the models all simulate the same climate response, then the only source of uncertainty in the ensemble expected climate response β_F is the internal variability. Therefore, the confidence interval for β_F based on the two-way framework of Equation 3.5 contains all of the uncertainty about the expected climate response of the ensemble. However, we must still assume that any discrepancy between the expected climate of the ensemble and the actual climate is constant in order to justify $\hat{\beta}_F$ as an estimate of the actual climate response.

3.3.3. A one-way ANOVA framework

Climate models rarely reproduce the observed climate accurately, they exhibit a variety of biases and deviations from the historical record. In the unlikely event that the climate models all simulate the same historical climate (i.e., $\alpha_m = 0 \; \forall \; m$), then estimating the α_m effects is unnecessary. In that case, an even simpler one-way ANOVA framework should provide more precise estimates

$$\begin{aligned} x_{smr} &= \mu + \beta_s + \varepsilon_{smr} \\ \varepsilon_{smr} &\overset{iid}{\sim} N\left(0, \sigma^2\right) \end{aligned} \tag{3.7}$$

with the constraint that $\beta_H = 0$. The effects have the same interpretation as in Equation 3.4. In Appendix A.3 it is shown that the maximum likelihood estimate of the ensemble expected climate response $\hat{\beta}_F$ from the one-way framework is a weighted average of the model mean responses with weights

$$w_{Hm} = N_{Hm} \quad \text{and} \quad w_{Fm} = N_{Fm} \tag{3.8}$$

In this case, equal weight is given to each run in the ensemble. This makes sense if all models really do simulate the same historical climate and climate response. Only $P = 2$ parameters must be estimated for the one-way framework. Therefore the

precision should increase since there are $M-1$ less parameters to estimate compared to the two-way framework and so an additional $M-1$ degrees of freedom available to estimate the internal variability. However, if the models do not all simulate the same historical climate, then a systematic component is missing from the framework and the estimates will be biased (Appendix A.4). Therefore, the one-way framework should only be used when we are confident that the models all simulate the same climate response *and* the same historical climate. Under the assumption that any discrepancy between the expected climate of the models and the actual climate is constant between scenarios, then $\hat{\beta}_F$ is still an unbiased estimate of the actual climate response. However, $\hat{\mu}$ will not be unbiased for the actual historical climate unless there is no discrepancy in the historical period.

3.4. Assumptions and framework checking

The goal of constructing a statistical framework to describe the ensemble is to make inferences about the values of key parameters, e.g., the expected climate response β_F. If we are to have confidence in those inferences, then the assumptions underlying the statistical framework must be understood and checked. In this section, the ANOVA frameworks are used to make explicit one set of assumptions that lead naturally to the "one model, one vote" estimate of the climate response. Simple graphical techniques and hypothesis tests are outlined so that those assumptions can be checked. If the assumptions are satisfied then we should have good confidence in inferences based on the ANOVA frameworks.

3.4.1. Assumptions

Since no attempt is made to weight the models according to any performance metric, it is assumed that all the climate models are equally valid representations of the Earth system. Performance criteria or expert judgement could be used to select a subset of the models for analysis, prior to fitting a statistical framework. If performance criteria are used, then ideally several metrics should be compared, as models will perform differently depending on the measure used for comparison (Gleckler et al., 2008).

ANOVA frameworks are usually estimated based on simple linear combinations of the group means over the factors included in the framework, in this case the model-scenario means $\overline{x}_{sm\cdot}$. In order to achieve that simplicity, a balanced design is required so that each model makes the same number of runs of each scenario. This restriction can be overcome by fitting the ANOVA framework using normal linear regression

methods (Krzanowski, 1998). The remaining assumptions are the standard assumptions of linear regression, specifically that:

- the ε_{smr} are normally distributed,
- the ε_{smr} have zero mean and constant variance,
- the ε_{smr} are mutually independent.

The ε_{smr} terms represent departures from the expected climate of a model due to unforced internal variability. Provided that the climate response trend is small and that individual years can be considered roughly independent, the central limit theorem implies that the uncertainty about any long term mean (e.g., a 30-year climate normal) will be approximately normally distributed. So, the assumption of normality is expected to hold for a wide range of climate variables.

The assumption of constant variance implies that the magnitude of the internal variability is constant for all models and both scenarios. Climate models are known to simulate different internal variability (Buser et al., 2009; Hawkins and Sutton, 2009). However, most of the CMIP5 models do not have sufficient runs to permit precise estimation of their individual variabilities. This is most often the case for the future scenarios where some models only simulate one initial condition run. The simplifying assumption of constant variance allows us to borrow strength across models *and* scenarios in order to estimate the internal variability.

It is assumed that the mean climates of runs from the same model and scenario are independent. Over short time scales, perhaps as long as months to a few years, there may be some correlation due to low frequency variability in the climate system (Collins et al., 2006b). However, over longer time scales ensemble members will diverge and evolve independently (Deser et al., 2012a). The assumption of independence also implies that the expected climate response of each model is independent of its expected historical climate. While this is usually assumed to be the case, it is not always the case (e.g., Bracegirdle and Stephenson, 2012) and therefore should be checked.

Finally, the assumption of independence also implies that the mean climates of the individual models are independent. There is empirical evidence that model biases from the actual climate are not independent (Knutti et al., 2010b). However, the model departures from the ensemble mean may be independent (Chandler, 2013; Rougier et al., 2013). This assumption may still be optimistic since models often share numerical codes or even entire components. However, it less strong than the increasingly unsupportable assumption of independence relative to the actual climate and therefore more acceptable. If the models are not independent, then there is less information in the ensemble than the number of models / runs would

suggest. In such situations, our inferences may be overconfident. In Chapter 4, we consider thinning the ensemble in order to reduce the effect of dependence between similar models.

3.4.2. Framework checking

The assumption of normality can be checked by plotting the ordered standardised residuals e' (Equation A.8) from the ANOVA framework against the theoretical quantiles of the normal distribution. If the runs are normally distributed, then the points should lie close to a straight line through the origin with unit gradient. If estimates are required for a large number grid points, then the Anderson-Darling test can be used to formally test the hypothesis of normality. While not as sensitive as the quantile-quantile plot, the Anderson-Darling test has greater power than the more general Kolmogorov-Smirnoff test against a range of departures from normality (Stephens, 1974).

In principle, the assumption of constant variance between models could be checked by splitting long control runs into series of non-overlapping 30 year periods. However, significance tests for non-constant variance lack power in small samples (Brown and Forsythe, 1974) and this technique cannot be applied to transient future scenarios. Without *very* long control runs, it is unlikely that any formal test will detect a difference between models unless the difference is large. In such situations, simple graphical checks are sufficient to detect any problems. Plotting the standardised residuals from the linear regression against the fitted values should show a random scatter about the zero line if the runs are symmetrically distributed with constant variance. Any differences in the internal variability between the historical and future scenarios should also be visible in this plot. Identifying runs from the different models / scenarios using different symbols or colours will highlight any specific violations.

The assumption of independence between the climate response and historical climate can be checked by plotting the estimates of the γ_{Fm} terms against the estimates of the α_m terms, i.e., the model specific components of the climate response against the model specific components of the historical climate. If the response is independent of the historical state, then the points will be randomly scattered about the origin. If the assumption needs to be checked for a large number of grid points, then plotting the correlation between the estimated γ_{Fm} and α_m terms will identify regions with strong linear dependence.

Dependence between models has been empirically investigated by looking at the distribution of correlations between all possible pairs of maps of biases between

models and observations (Knutti et al., 2010b). If the models are independently distributed about the ensemble mean rather than the actual climate, then the correlations between all possible maps of the model specific departures α_m or γ_{Fm} should be distributed around zero. A slight negative correlation is actually expected since each model contributes to the sample estimate of the ensemble mean.

3.4.3. Identifying influential ensemble members

In Section 3.2, it was noted that one of the problems with the "one model, one vote" approach to climate projection is that arithmetic means are not robust estimators. Runs that behave very differently to the rest of the ensemble may strongly affect the estimated climate response. Such runs should not necessarily be removed from the ensemble, since they may represent perfectly plausible climates, simply ones that lie in the tails of the distribution of possible climates. They may contribute valuable information to the ensemble. On the other hand, there are a variety of circumstances that may lead to undesirable results entering the ensemble, e.g., poorly chosen initial conditions, user error at setup, an inflexible process parameterisation, or errors arising in post-processing.

The distributional assumptions of the linear frameworks can be used to identify potentially problematic runs. Runs with large standardised residuals, i.e., $|e'_{smr}| \gg t_{N_{\cdot\cdot}-P}\,(0.995) \approx 2.5$, may be regarded as outlying compared to the rest of the ensemble and flagged for further investigation. During the preparation of Sansom et al. (2013), this simple check identified a run of the MIROC5 model which behaved very differently to the rest of the ensemble. Further investigation revealed that an error had occurred during the post-processing of the storm track data.

A run may also be influential through the values of the explanatory variables used in the linear regression. Such runs are described as having high leverage. In the frameworks described in Section 3.3, the explanatory variables are binary indicators of the model and scenario which made the run. Therefore, all runs from the same model and scenario will have the same leverage. Particular attention should be paid to outlying runs from model-scenario pairs with high leverage. The uncertainty associated with the fitted value $\hat{x}_{smr}$ of a data point in a linear regression framework is directly proportional to its leverage (Krzanowski, 1998, Section 3.5). Therefore, the leverage of each model-scenario pair under the three ANOVA frameworks is proportional to Equations A.5, A.12 & A.18 (Appendix A). Under the framework with interactions, the leverage of a run depends on the number of runs of that scenario by that model N_{sm}. Outlying runs from model-scenario pairs with only one initial condition run may be particularly influential. Under the two-way framework, the leverage is highest for models with only a few initial condition runs in total. The

leverage under the one-way framework depends only on the total number of runs of that particular scenario (from all models). Therefore outlying runs from a scenario with very few initial condition runs may be particularly influential.

One way to check the influence of a particular run is to remove it from the ensemble, refit the statistical framework and compare the estimates of the climate response β_F. If the estimates are similar, then the influence is small and the run can safely remain in the ensemble. If the estimates are very different then it is important to determine whether there really is a problem with the run *before* discarding it from the ensemble.

Since the estimates of the expected climate response β_F derived in Section 3.3 are all weighted averages of the model mean climates, the weights determine the influence of a particular model on the estimate of the expected climate response. Therefore, if a particular model has both a high weight (w_{Hm} or w_{Fm}) and a large departure (α_m or γ_{Fm}) for either scenario, it will have a large effect on the expected climate response. Under the two-way framework with interactions, equal weight is given to each scenario and each model. So the most influential models will be those with the largest response departures γ_{Fm}. The estimated response departures from the framework with interactions are also a useful indicator of influence for the two-way framework, since it also gives equal weight to both scenarios. However models with a large number of run receive additional weight, especially those with many runs of *both* scenarios. Under the one-way framework, a model with many runs and a large departure in either scenario will be influential. If a particular model is suspected of strongly influencing the estimate $\hat{\beta}_F$, then it could be removed, the framework refitted, and the estimates compared. However, as with individual runs, an influential model should not be discarded simply because it is influential. It is important to determine whether the model really is unrealistic in some way, or whether it is contributing valuable information. It could be the case that the influential model is the one that most resembles the observed climate.

3.5. Inference in the linear regression frameworks

In this section, the inferential tools of linear regression are applied and interpreted within the frameworks described in Section 3.3. In particular, it is shown that F tests can be used to test for climate model agreement, and t tests can be used to test for evidence of a non-zero climate response.

3.5.1. Do all the models simulate the same climate response?

In Section 3.3.2, it was noted that the two-way framework should only be used when we are confident that all models simulate the same climate response. Agreement between models is usually quantified by the number of models agreeing on the sign of the mean response (e.g., Alley and Coauthors, 2007, Figure SPM.7). This approach fails to account for the effect of internal variability, particularly when only one run of each model is used. If the climate change response is small compared to the size of the internal variability (i.e., $\beta_F \ll \sigma$), then even if the models all simulate the same climate response, they will not agree on the sign.

In the framework with interactions, agreement on the expected climate response corresponds to the condition that $\gamma_{Fm} = 0 \; \forall \; m = 1, \ldots, M$. The null hypothesis $H_0 : \gamma_{Fm} = 0 \; \forall \; m = 1, \ldots, M$ can be compared to the alternative hypothesis $H_\gamma : \gamma_{Fm} \neq 0$ for some m by comparing the likelihoods under each hypothesis, i.e., by comparing the likelihoods of the two-way framework and the framework with interactions. A suitable test statistic (Davison, 2003) is the standardised variance ratio

$$F_\gamma = \frac{N_{\cdot\cdot} - 2M}{M - 1} f_\gamma^2 \quad \text{where} \quad f_\gamma^2 = \frac{R_\gamma^2 - R_\alpha^2}{1 - R_\gamma^2} \tag{3.9}$$

and R_γ^2 and R_α^2 are the coefficients of determination of the framework with interactions and the two-way framework respectively. The coefficient of determination is the proportion of the total variability in the ensemble explained by the regression framework. Therefore, the quantity f_γ^2 represents the ratio of the variance explained by the inclusion of the γ_{Fm} terms, i.e., the variance due to model differences in the climate response, compared to the variance explained by internal variability. So model agreement corresponds to the model uncertainty in the response being small compared to the internal variability. Under the null hypothesis H_0, F_γ has a F distribution with $M - 1$ and $N_{\cdot\cdot} - 2M$ degrees of freedom. For a test of size $100\alpha\%$, we should reject the null hypothesis of model agreement when F_γ exceeds the $1 - \alpha$ quantile of the F distribution.

Like all hypothesis tests, a result that is not significant does not mean that the null hypothesis is true, only that there is insufficient evidence to rule it out. A more informative approach would be to exploit the duality between hypothesis tests and confidence intervals in order to construct an interval estimate for the level of agreement between the models, but how should such an interval be defined? The standard F distribution arises as the ratio of two χ^2 distributions. The standard χ^2 distribution is a sum of squared *standard* normal deviates. Under the the alternative hypothesis H_γ, the numerator of f_γ^2 in Equation 3.9 has a non-central χ^2 distribution

(Johnson and Kotz, 1970, Chapter 28) since the expected values of at least some of the response departures γ_{Fm} are not 0. Therefore, the test statistic F_γ has a non-central F distribution (Johnson and Kotz, 1970, Chapter 30) with non-centrality parameter

$$\lambda_\gamma = \sum_{m=1}^{M} \sum_{s\in\{H,F\}} \sum_{r=1}^{N_{sm}} \left(\frac{\gamma_{Fm}}{2\sigma}\right)^2 = \frac{1}{4}\sum_{m=1}^{M} N_{\cdot m} \left(\frac{\gamma_{Fm}}{\sigma}\right)^2 \tag{3.10}$$

where $N_{\cdot m} = N_{Hm} + N_{Fm}$ is the total number of runs from model m. The factor of $1/2$ arises due to the fact that the interaction effects were defined using "treatment" contrasts $\gamma_{Hm} = 0 \;\forall\; m$ and $\sum_{m=1}^{M} \gamma_{Fm} = 0$, rather than standard "sum to zero" contrasts $\sum_{m=1}^{M} \gamma_{sm} = 0 \;\forall\; s$ and $\sum_{s\in\{H,F\}} \gamma_{sm} = 0 \;\forall\; m$. From Equation 3.10 we see that λ_γ depends on both the number of models and the total number of runs from each model. However, for a balanced ensemble where $N_{\cdot m} = N \;\forall\; m$, we can define the following relationship (Steiger, 2004)

$$\Psi_\gamma = \sqrt{\frac{2\lambda_\gamma}{N(M-1)}} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\frac{\gamma_{Fm}}{\sigma}\right)^2} \tag{3.11}$$

The quantity Ψ_γ is easily interpreted as the standardised root-mean-square of the inter-model spread in the climate response. The factor $M-1$ appears due to the constraint that the model departures sum to zero, so there are only $M-1$ independent parameters. The usual measure of model agreement, the number of models that agree on the sign of the climate response, can also be interpreted as a function of the inter-model spread. The relationship defined above shows that is important to consider the spread of the ensemble in relation to the size of internal variability. The uniformly minimum variance unbiased estimate of λ_γ is (Johnson et al., 1995)

$$\hat{\lambda}_\gamma = f_\gamma^2 \left(\nu_2 - 2\right) - \nu_1 \tag{3.12}$$

where $\nu_1 = M-1$ and $\nu_2 = N_{\cdot\cdot} - 2M$ are the degrees of freedom associated with F_γ. A $100(1-\alpha)\%$ confidence interval for λ_γ is given by $\left(\lambda_\gamma^l \;,\; \lambda_\gamma^u\right)$ where

$$\lambda_\gamma^l = \inf\{\lambda : \Pr\left(X \le F_\gamma \,|X \sim F_{\nu_1,\nu_2,\lambda}\right) \ge 1-\alpha/2\} \tag{3.13a}$$

$$\lambda_\gamma^u = \sup\{\lambda : \Pr\left(X \le F_\gamma \,|X \sim F_{\nu_1,\nu_2,\lambda}\right) \le \alpha/2\} \tag{3.13b}$$

This interval can only be evaluated numerically, by iteratively calculating the percentile of the non-central F distribution corresponding to the observed value of F_γ for particular values of λ. Since Ψ_γ is a monotonic increasing function of λ_γ, a confidence interval for Ψ_γ can be obtained by transforming the confidence interval for λ_γ using Equation 3.11. If the lower limit Ψ_γ^l is *not* 0, then we should reject the null hypothesis of model agreement at the $100\alpha\%$ level. However, we can also evaluate the upper limit Ψ_γ^u in order to assess whether the size of the structural

uncertainty in the model response is likely to be below an acceptable level. If the upper limit Ψ_γ^u is much smaller than the standardised climate response $\left|\hat{\beta}_F/\sigma\right|$, then we might be willing to conclude that the models agree sufficiently to make useful inferences, even if the F test rejects the null hypothesis of model agreement. If we suppose that the response departures γ_{Fm} arise from a normal distribution centred on β_F, then 95% of the model mean responses would lie within approximately $2\,\mathrm{RMS}\,(\gamma_{Fm})$ of β_F. Therefore, we should require that $\Psi_\gamma^u < \left|\hat{\beta}_F/2\sigma\right|$ or $\mathrm{RMS}\,(\hat{\gamma}_{Fm}) < \left|\hat{\beta}_F/2\right|$ as a minimal condition for model agreement. Alternatively, $\Psi_\gamma^u < |\beta_F^*/2\sigma|$ or $\mathrm{RMS}\,(\hat{\gamma}_{Fm}) < |\beta_F^*/2|$ if we are interested in a particular climate response β_F^*.

3.5.2. Do all the models simulate the same historical climate?

A similar analysis can be performed for agreement between models on the expected historical climate. Under the two-way framework, agreement on the historical climate implies that $\alpha_m = 0\ \forall\ m = 1, \ldots, M$. The null hypothesis $H_0 : \alpha_m = 0\ \forall\ m = 1, \ldots, M$ can be compared to the alternative hypothesis $H_\alpha : \alpha_m \neq 0$ for some m using the standardised variance ratio

$$F_\alpha = \frac{N_{\cdot\cdot} - (M+1)}{M-1} f_\alpha^2 \quad \text{where} \quad f_\alpha^2 = \frac{R_\alpha^2 - R_\beta^2}{1 - R_\alpha^2} \tag{3.14}$$

and R_β^2 is the coefficient of determination of the one-way framework. The dimensionless ratio f_α^2 represents the ratio of the variance explained by the inclusion of the α_m terms, i.e., the variance due to model uncertainty in the historical climate, to the variance explained by internal variability. So model agreement on the historical climate corresponds to the model uncertainty in the historical climate being small compared to the internal variability. Under the null hypothesis H_0, F_α has a F distribution with $M-1$ and $N_{\cdot\cdot} - (M+1)$ degrees of freedom. For a test of size $100\alpha\%$, we should reject the null hypothesis of model agreement when F_α exceeds the $1-\alpha$ quantile of the F distribution.

Under the alternative hypothesis H_α, the test statistic F_α has a non-central F distribution with non-centrality parameter

$$\lambda_\alpha = \sum_{m=1}^{M} \sum_{s \in \{H,F\}} \sum_{r=1}^{N_{sm}} \left(\frac{\alpha_m}{\sigma}\right)^2 \tag{3.15}$$

Like λ_γ, the non-centrality parameter λ_α depends on both the number of models and the number of runs from each model. So, for a balanced ensemble where $N_{sm} =$

$N \forall\, s, m$, define

$$\Psi_\gamma = \sqrt{\frac{\lambda_\alpha}{2R(M-1)}} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\frac{\alpha_m}{\sigma}\right)^2} \tag{3.16}$$

The statistic Ψ_α is interpreted as the standardised root-mean-square of the inter-model spread in the historical climate. The non-centrality parameter can be estimated by

$$\hat{\lambda}_\alpha = f_\alpha^2\,(\nu_2 - 2) - \nu_1 \tag{3.17}$$

where $\nu_1 = M - 1$ and $\nu_2 = N_{\cdot\cdot} - (M+1)$ are the degrees of freedom associated with F_α. A $100(1-\alpha)\%$ confidence interval for λ_α is given by $\left(\lambda_\alpha^l\ ,\ \lambda_\alpha^u\right)$ where

$$\lambda_\alpha^l = \inf\{\lambda : \Pr\left(X \leq F_\alpha\,|X \sim F_{\nu_1,\nu_2,\lambda}\right) \geq 1 - \alpha/2\} \tag{3.18a}$$

$$\lambda_\alpha^u = \sup\{\lambda : \Pr\left(X \leq F_\alpha\,|X \sim F_{\nu_1,\nu_2,\lambda}\right) \leq \alpha/2\} \tag{3.18b}$$

A confidence interval for Ψ_α can be constructed by transforming the interval for λ_α using Equation 3.16. If the lower limit Ψ_α^l is *not* zero then we should reject the null hypothesis of model agreement at the $100\alpha\%$ level. For the climate response,a minimal condition for model agreement was proposed based on the upper limit of the confidence interval Ψ_γ^u. In Appendix A.4, it is shown that under the one-way framework any historical differences will bias the estimate of β_F. In order to keep the contribution to the bias from each model small, it is sensible to adopt the same upper bound as for the model responses. Therefore, we should require that $\Psi_\alpha^u < \left|\hat{\beta}_F/2\right|$, or equivalently $\mathrm{RMS}\left(\hat{\alpha}_m\right) < \left|\hat{\beta}_F/2\right|$, as a minimal condition for model agreement. Alternatively, $\Psi_\alpha^u < |\beta_F^*/2|$ or $\mathrm{RMS}\left(\hat{\alpha}_m\right) < |\beta_F^*/2|$ if we are interested in a particular climate response β_F^*.

3.5.3. Is there evidence of a climate response?

The parameter estimates in the linear regression frameworks are all linear combinations of the climates of the individual runs. The runs are assumed to be normally distributed, and linear combinations of normal random variables are also normally distributed. However, the internal variability σ^2 is unknown and must also be estimated (Equation A.2). Therefore, the parameter estimates will have t distributions with $N_{\cdot\cdot} - P$ degrees of freedom. The null hypothesis of no climate response ($H_0 : \beta_F = 0$) can be compared against the alternative hypothesis of any climate

response ($H_\beta : \beta_F \neq 0$) using the test statistic

$$T_\beta = \left| \frac{\hat{\beta}_F - 0}{\sqrt{\text{var}\left(\hat{\beta}_F\right)}} \right| \tag{3.19}$$

Under the null hypothesis, T_β has a standard t distribution with $N_{\cdot\cdot} - P$ degrees of freedom. H_β is a two-sided alternative, so for a test of size $100\alpha\%$, we should reject the null hypothesis of no climate response if T_β exceeds the $1 - \alpha/2$ quantile of the t distribution.

A more informative approach is to construct a confidence interval for β_F

$$\hat{\beta}_F - t_\nu\left(\alpha/2\right)\sqrt{\text{var}\left(\hat{\beta}_F\right)} \;\leq\; \beta_F \;\leq\; \hat{\beta}_F + t_\nu\left(1-\alpha/2\right)\sqrt{\text{var}\left(\hat{\beta}_F\right)} \tag{3.20}$$

where $t_\nu\left(\alpha\right)$ is the quantile function of the t distribution with ν degrees of freedom. If 0 does *not* lie within the confidence interval, then the null hypothesis of no climate response is rejected at the $100\alpha\%$ level. However, we are now free to consider whether any particular climate response is supported by the ensemble. Suppose that a response greater than β_F^* would have a dangerous impact in a particular region. If the upper limit of the confidence interval for β_F is less than β_F^*, then the ensemble provides no evidence at the $100\alpha\%$ level that we need to be concerned about such a scenario occurring.

Note that if the models do not agree on the climate response, then the confidence interval for β_F will not include the additional uncertainty due to model uncertainty. Therefore, we should only make inferences about the expected climate response if we are satisfied that the models agree sufficiently to support them, as described in Section 3.5.1.

The measures of model agreement defined in the previous two sections are based on standardised measures of model departures. It is therefore also useful to define the standardised climate response or signal-to-noise ratio of the climate response

$$d_\beta = \frac{\beta_F}{\sigma} \tag{3.21}$$

Like the standardised RMS measures defined for the F tests, this is easily understood on the scale of internal variability, i.e., $d_\beta > 2$ indicates the ensemble expected future climate is more extreme than 95% of plausible historical climates. Confidence intervals for the standardised climate response can be constructed in the same way as for the other standardised measures, by reference to the distribution of the test statistic T_β under the alternative hypothesis. If $\beta_F \neq 0$, then T_β has a non-central t

distribution (Johnson and Kotz, 1970, Chapter 27) with $N_{\cdot\cdot} - P$ degrees of freedom and non-centrality parameter

$$\mu_\beta = \frac{\beta_F}{\sqrt{\text{var}\,(\beta_F)}} = \frac{\beta_F}{\sigma/\sqrt{k}} = d_\beta\sqrt{k} \tag{3.22}$$

where k is a constant that depends only on the number of models and initial condition runs, and which framework is fitted (see Appendix A, Equations A.4c, A.11c and A.17b). A $100(1-\alpha)\%$ confidence interval for μ_β is given by

$$\mu_\beta^l = \inf\{\mu : \Pr\,(X \leq T_\beta | X \sim t_{\nu,\mu}) \geq 1 - \alpha/2\} \tag{3.23a}$$

$$\mu_\beta^u = \sup\{\mu : \Pr\,(X \leq T_\beta | X \sim t_{\nu,\mu}) \leq \alpha/2\} \tag{3.23b}$$

where $\nu = N_{\cdot\cdot} - P$ is the degrees of freedom associated with T_β. Like the intervals for the standardised RMS measures, this can only be evaluated numerically. Since μ_β is a monotonic function of d_β for fixed k, a confidence interval for d_β can be obtained by transforming the interval for μ_β using Equation 3.22. Note that this interval will be different to one obtained by simply dividing through Equation 3.20 by σ. That interval would fail to take into account the uncertainty in the value of σ. The interval based on Equation 3.23 correctly accounts for the uncertainty about σ and will therefore be non-symmetric.

3.5.4. Do the individual model responses agree with the expected climate response?

A similar t test can be used to test the hypothesis that an individual model agrees with the expected response of the ensemble. The null hypothesis of no departure for model m ($H_0 : \gamma_{Fm} = 0$) can be compared with the alternative hypothesis of any departure for model m ($H_\gamma : \gamma_{Fm} \neq 0$) using the test statistic

$$T_\gamma = \left| \frac{\hat{\gamma}_{Fm} - 0}{\sqrt{\text{var}\,(\hat{\gamma}_{Fm})}} \right| \tag{3.24}$$

Under the null hypothesis, T_γ has a standard t distribution with $N_{\cdot\cdot} - P$ degrees of freedom. H_γ is a two-sided alternative, so for a test of size $100\alpha\%$, we should reject the null hypothesis of no climate response if T_γ exceeds the $1 - \alpha/2$ quantile of the t distribution. If required, confidence intervals for the true value of γ_{Fm} can be constructed in a manner analogous to Equation 3.20. This test is useful in combination with the criteria outlined at the end of Section 3.4.3 to identify models with large departures that may strongly influence the estimate of the climate response. Particular attention should be paid to models with significant departures

and only a small number of initial condition runs.

A similar test is available to test whether an individual model agrees with the expected historical climate of the ensemble. The null hypothesis of no departure for model m ($H_0 : \alpha_m = 0$) can be compared with the alternative hypothesis of any departure for model m ($H_\alpha : \alpha_m \neq 0$) using the test statistic

$$T_\gamma = \left| \frac{\hat{\alpha}_m - 0}{\sqrt{\text{var}\left(\hat{\alpha}_m\right)}} \right| \tag{3.25}$$

Under the null hypothesis, T_α also has a standard t distribution with $N_{\cdot\cdot} - P$ degrees of freedom. H_α is a two-sided alternative, so for a test of size $100\alpha\%$, we should reject the null hypothesis of no climate response if T_α exceeds the $1 - \alpha/2$ quantile of the t distribution.

Once again, it is important to emphasise that models that do not agree with the expected climate or climate response should not be removed from the ensemble simply because they do not agree. It is useful to be able to systematically identify such models for further investigation. However, they should only be removed if either detailed investigation or expert judgement indicate that doing so is warranted.

3.6. Is the ensemble large enough?

In Section 3.5, a new measure of climate model agreement was described along with inferential procedures for evaluating the evidence of a climate response. The precision of the estimates (the width of the confidence intervals) depends on the number of models and runs in the ensemble. How large would an ensemble need to be for our estimates to be sufficiently precise so as to always detect a climate response of a particular size, or model disagreement over a certain threshold? In order to answer this question we need to consider the power of the hypothesis tests. The power of a test is the probability of rejecting the null hypothesis given that an effect really exists, and it depends on the size of the test, the size of the ensemble, and the size of the effect we wish to detect.

3.6.1. Power of *t* tests

The distribution of the test statistic T_β under the alternative hypothesis $H_\beta : \beta_F \neq 0$ was already defined in Section 3.5.4. There it was shown that the non-centrality parameter μ_β which defines the distribution under H_β could be written in terms of the standardised climate response d_β. Given the value of the non-centrality parameter,

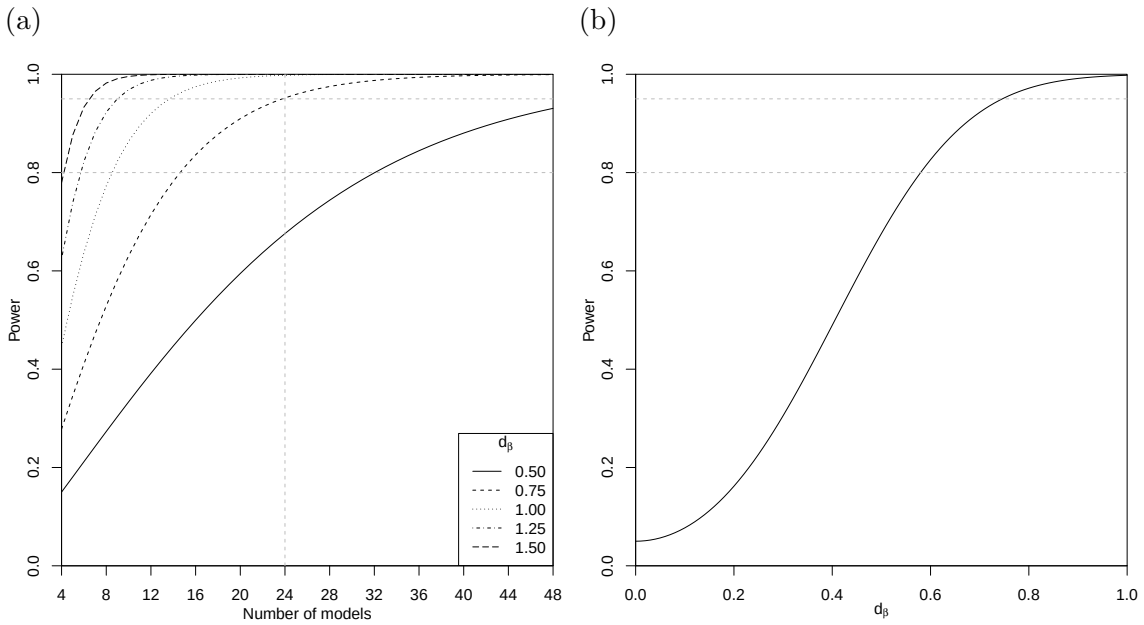

Figure 3.1.: (a) Power of the t test for non-zero climate response as a function of ensemble size, for various standardised climate responses d_β, based on two runs of each scenario from each model. The dashed grey vertical line indicates an ensemble with 24 models, similar to the CMIP5 ensemble analysed in this thesis; (b) Power of the t test as a function of the standardised climate response d_β for an ensemble of similar size to CMIP5 ensemble analysed in this thesis, with 24 models and two runs of each scenario. Dashed horizontal grey lines indicate the 80% and 95% power levels.

the power of the t test for non-zero climate response is

$$\Pr\left(T_\beta \leq t_\nu\left(\alpha/2\right) \mid T_\beta \sim t_{\nu,\mu_\beta}\right) + \Pr\left(T_\beta > t_\nu\left(1-\alpha/2\right) \mid T_\beta \sim t_{\nu,\mu_\beta}\right) \tag{3.26}$$

where $\nu = N_{\cdot\cdot} - P$ is the degrees of freedom associated with the estimate $\hat{\beta}_F$, $t_\nu\left(\alpha\right)$ is the quantile function of the central t distribution with ν degrees of freedom. The power of the t test is shown for various ensemble sizes and standardised climate responses in Figure 3.1a. For an ensemble similar to CMIP3 or the subset of CMIP5 models analysed in this thesis (Figure 3.1b), the power exceeds 95% for $d_\beta \geq 0.75$ and exceeds 80% for $d_\beta \geq 0.58$, or equivalently $\beta_F \geq 0.75\sigma$ and 80% for $\beta_F \geq 0.58\sigma$ respectively. The full CMIP5 ensemble used in the IPCC Fifth Assessment Report (Collins et al., 2013), contains 45 models, and will be able to detect even smaller changes, assuming that the additional models provide independent information.

The results in Figure 3.1 show that the CMIP5 ensemble analysed in this thesis is large enough to reliably detect relatively small climate responses. If we want to be confident of detecting even smaller responses then we either have to increase the ensemble size, or increase the test size. Increasing the test size increases the risk of rejecting the null hypothesis when there really is no response. When the potential impact is high, then this may be acceptable, however this must be balanced against

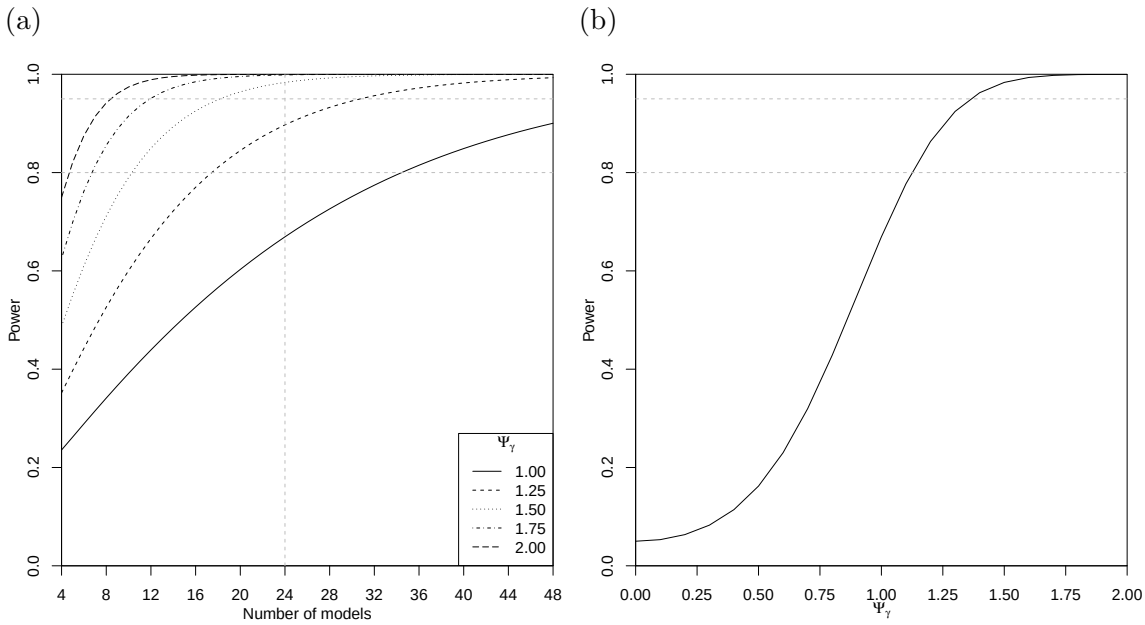

Figure 3.2.: (a) Power of the F test for model agreement on the climate response as a function of ensemble size, for various standardised root-mean-squares Ψ_γ, based on two runs of each scenario from each model. The dashed grey vertical line indicates an ensemble of similar size to the CMIP5 ensemble analysed in this thesis with 24 models; (b) Power of the F test as a function of the standardised root-mean-square Ψ_γ for an ensemble similar the CMIP5 ensemble analysed in this thesis with 24 models and two runs of each scenario. Dashed grey horizontal lines indicate the 80% and 95% power levels.

the potential costs of taking unnecessary action.

3.6.2. Power of F tests

The distribution of the statistic F_γ under the alternative hypothesis was already defined in Section 3.5. There it was shown that the non-centrality parameter (λ_γ) could be written in terms of the standardised RMS of the inter-model spread in the climate response (Ψ_γ). So, given the non-centrality parameter, the power of the F test for model agreement on the climate response is

$$\Pr\left(F_\gamma > F_{\nu_1,\nu_2}(\alpha) \mid F \sim F_{\nu_1,\nu_2,\lambda_\gamma}\right) \tag{3.27}$$

where $\nu_1 = M - 1$ and $\nu_2 = N_{\cdot\cdot} - 2M$ are the degrees of freedom associated with F_γ and $F_{\nu_1,\nu_2}(\alpha)$ is the quantile function of the central F distribution with ν_1 and ν_2 degrees of freedom. The power of the F test is shown for various ensemble sizes and values of Ψ_γ in Figure 3.2a. For an ensemble similar to that analysed in this thesis (Figure 3.2b), the probability of detecting model disagreement in the response exceeds 95% for $\Psi_\gamma \geq 1.39$ and 80% for $\Psi_\gamma \geq 1.15$, or equivalently $\mathrm{RMS}(\gamma_{Fm}) \geq 1.39\sigma$ and $\mathrm{RMS}(\gamma_{Fm}) \geq 1.15\sigma$ respectively. In contrast to the t tests, only quite

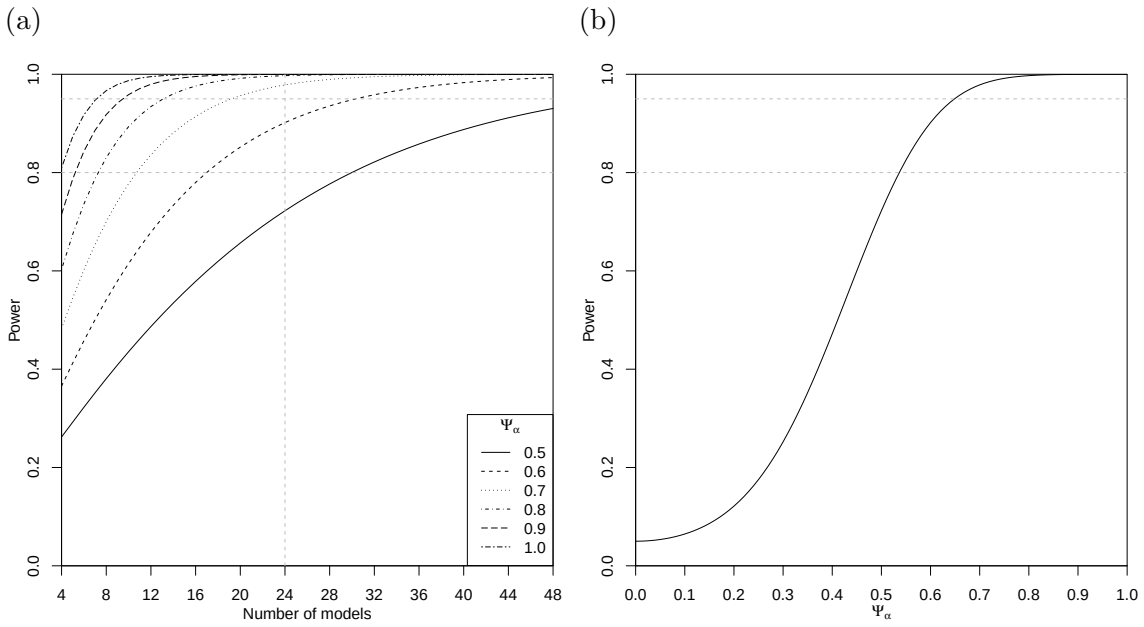

Figure 3.3.: (a) Power of the F test for model agreement on the historical climate as a function of ensemble size for various standardised root-mean-squares Ψ_α, based on two runs of each scenario from each model. The dashed grey vertical line indicates an ensemble similar to the CMIP5 ensemble analysed in this thesis with 24 models; (b) Power of the F test as a function of the standardised root-mean-squares Ψ_α for an ensemble similar to the CMIP5 ensemble analysed in this thesis with 24 models and two runs of each scenario. Dashed grey horizontal lines indicate the 80% and 95% power levels.

large differences between model mean responses can be reliably detected in the subset of the CMIP5 ensemble analysed in this thesis.

Similarly, given the non-centrality parameter (λ_α) defined in terms of the standardised RMS (Ψ_α), the power of the F test for model agreement on the historical climate is

$$\Pr\left(F_\alpha > F_{\nu_1,\nu_2}\left(\alpha\right) | F \sim F_{\nu_1,\nu_2,\lambda_\alpha}\right) \tag{3.28}$$

where $\nu_1 = M - 1$ and $\nu_2 = N_{\cdot\cdot} - (M + 1)$ are the degrees of freedom associated with F_α. The power of the F test is shown for various ensemble sizes and values of Ψ_α in Figure 3.3a. For an ensemble similar to the subset of CMIP5 models analysed in this thesis (Figure 3.3a), the probability of detecting model disagreement in the response exceeds 95% for $\Psi_\alpha \geq 0.67$ and 80% for $\Psi_\alpha \geq 0.55$, or equivalently $\mathrm{RMS}\left(\alpha_{Fm}\right) \geq 0.67\sigma$ and $\mathrm{RMS}\left(\alpha_{Fm}\right) \geq 0.55\sigma$ respectively. Much smaller differences between the historical climates of the CMIP5 models can be reliably detected than differences between their responses. This comparison is unfortunate since the differences between the responses tend to be small compared to the differences between the historical climates.

3.7. Framework selection strategy

The frameworks outlined in Section 3.3 form a nested hierarchy. The one-way framework is a special case of the two-way framework where $\alpha_m = 0 \; \forall \; m$. The two-way framework is also a special case of the framework with interactions where $\gamma_{Fm} = 0 \; \forall \; m$.

In Appendix A.4, it is shown that the framework with interactions provides an unbiased estimate of the ensemble expected climate response (β_F), even where we might prefer one of the simpler frameworks. Therefore, a simple approach to selecting the most appropriate framework is to compare the estimates $\hat{\beta}_F$ from the three frameworks. If the two-way estimate is similar to the estimate including interactions, then the two-way framework is probably sufficient to describe the ensemble. If the one-way estimate is also similar to the estimate including interactions, then the one-way framework is probably sufficient to describe the ensemble. Note that this strategy will fail in the case of a balanced ensemble, i.e., where all the models have the same number of runs ($N_{Hm} = N_H \; \forall \; m$ and $N_{Fm} = N_F \; \forall \; m$). In that case, the three estimates of β_F are equivalent and any bias will not be detectable.

The assumption checks and hypothesis tests outlined in Sections 3.4 and 3.5 provide a more rigorous approach that is able to distinguish between the estimates from the different frameworks even for a balanced ensemble:

1. Fit the framework with interactions.
2. Check the framework assumptions and identify any outlying runs.
    a) If the assumptions are satisfied and there are no outlying runs, then go to next step.
    b) If there are outlying runs, then remove them temporarily and repeat (2).
    c) If the assumptions are not satisfied and there are no outlying runs, then consider an alternative statistical framework or revert to the previous framework. If the assumptions of the framework with interactions are not satisfied, then a different framework or different distributional assumptions may be required.
3. Carry out the F test for model agreement on the climate response. If the null hypothesis of model agreement is rejected, then stop, the framework with interactions is most appropriate.
4. Fit the two-way framework.
5. Check the framework assumptions and identify and outlying runs as in (2).

6. Carry out the F test for model agreement on the historical climate. If the null hypothesis of model agreement is rejected, then stop, the two-way framework is most appropriate.

7. Fit the one-way framework.

8. Check the framework assumptions and identify and outlying runs as in (2).

This procedure is easily automated for large numbers of grid points with minimal user intervention. However, care must be taken when temporarily removing outliers to ensure that at least one run of each model is available for each scenario. Once the most appropriate framework has been selected, then the t test for non-zero climate response can be used to assess whether or not there is significant evidence of a climate response. Confidence intervals for the standardised effect sizes d_β, Ψ_γ and Ψ_α may also be useful for assessing the model agreement in comparison to the estimated response, or a particular response of interest.

3.8. Application to North Atlantic storm track

The ANOVA frameworks and inferential tools derived in this chapter are illustrated by applying them to the estimation of future climate change in the North Atlantic storm track, as simulated by 24 climate models participating in the fifth coupled model inter-comparison project (CMIP5) (Taylor et al., 2012). The variable of interest is the winter (December-January-February) extra-tropical cyclone track density, defined as the mean number of cyclones passing within 5° of a particular point each month. Cyclones are identified as maxima in the 850-hPa relative vorticity field and tracked through their life cycle using the TRACK algorithm (Hodges, 1994, 1995, 1999). Before tracking, the model output is filtered in order to remove large scale semi-permanent features of the background flow (Anderson et al., 2003; Hoskins and Hodges, 2002). The model output is also interpolated to a common resolution in order to reduce the noise in the vorticity fields and simplify comparisons between models (Hoskins and Hodges, 2002; Bengtsson et al., 2006). After tracking, only cyclones with a lifetime greater than 2 days and that travel at least 1000 km are retained. All cyclones are retained, not only strong cyclones, so not all of the systems included would be classed as storms. The track density is computed using the spherical kernel approach developed by Hodges (1996).

The data analysed here are an expanded version of those analysed in the studies of Sansom et al. (2013) and Zappa et al. (2013b). The study region (75W-45E, 30N-75N) is chosen to coincide with those earlier studies and covers the North Atlantic storm track and its exit region over Europe, as well as the Mediterranean storm

track and its exit region over the Middle East. The data represent monthly track density averaged over 30 winters from each of two time periods. The recent climate is represented by 30 winters between 1976 and 2005 from the CMIP5 historical scenario. The future climate is represented by 30 winters between 2070 and 2099 from the RCP4.5 medium mitigation scenario (Moss et al., 2010). A total of $N_{\cdot\cdot} = 98$ runs from $M = 24$ CMIP5 models are included, $N_{H\cdot} = 59$ from the historical scenario and $N_{F\cdot} = 39$ from the RCP4.5 scenario. The climate models included and the number of runs from each are listed in Table 3.1.

Table 3.1.: Number of realisations available from each model for the historical and future scenarios and the weights given by each linear regression framework. Weights have been standardised to sum to 100 for each framework.

Model	Runs		Weights					
	Historical	RCP4.5	Interactions		Two-way		One-way	
	N_{Hm}	N_{Fm}	w_{Hm}	w_{Fm}	w_{Hm}	w_{Fm}	w_{Hm}	w_{Fm}
BCC-CSM1.1	3	1	2.08	2.08	1.74	1.74	3.06	1.02
BCC-CSM1.1(m)	1	1	2.08	2.08	1.16	1.16	1.02	1.02
CanESM2	5	1	2.08	2.08	1.94	1.94	5.10	1.02
CCSM4	1	1	2.08	2.08	1.16	1.16	1.02	1.02
CMCC-CM	1	1	2.08	2.08	1.16	1.16	1.02	1.02
CNRM-CM5	5	1	2.08	2.08	1.94	1.94	5.10	1.02
CSIRO-Mk3.6.0	4	5	2.08	2.08	5.16	5.16	4.08	5.10
EC-EARTH	3	3	2.08	2.08	3.49	3.49	3.06	3.06
FGOALS-g2	1	1	2.08	2.08	1.16	1.16	1.02	1.02
GFDL-ESM2G	1	1	2.08	2.08	1.16	1.16	1.02	1.02
GFDL-ESM2M	1	1	2.08	2.08	1.16	1.16	1.02	1.02
HadGEM2-ES	1	1	2.08	2.08	1.16	1.16	1.02	1.02
HadGEM2-CC	2	1	2.08	2.08	1.55	1.55	2.04	1.02
INM-CM4	1	1	2.08	2.08	1.16	1.16	1.02	1.02
IPSL-CM5A-LR	4	4	2.08	2.08	4.65	4.65	4.08	4.08
IPSL-CM5A-MR	1	1	2.08	2.08	1.16	1.16	1.02	1.02
IPSL-CM5B-LR	1	1	2.08	2.08	1.16	1.16	1.02	1.02
MIROC5	5	3	2.08	2.08	4.36	4.36	5.10	3.06
MIROC-ESM	3	1	2.08	2.08	1.74	1.74	3.06	1.02
MIROC-ESM-CHEM	1	1	2.08	2.08	1.16	1.16	1.02	1.02
MPI-ESM-LR	3	3	2.08	2.08	3.49	3.49	3.06	3.06
MPI-ESM-MR	3	3	2.08	2.08	3.49	3.49	3.06	3.06
MRI-CGCM3	5	1	2.08	2.08	1.94	1.94	5.10	1.02
NorESM1-M	3	1	2.08	2.08	1.74	1.74	3.06	1.02
Total	59	39	50.00	50.00	50.00	50.00	60.20	39.80

3.8.1. The simple approach to framework selection

The left hand column of Figure 3.4 compares the usual multi-model mean estimate of the historical track density to the estimate from the ERA-Interim reanalysis (Dee et al., 2011). The mean storm track of the ensemble is too active near Newfoundland,

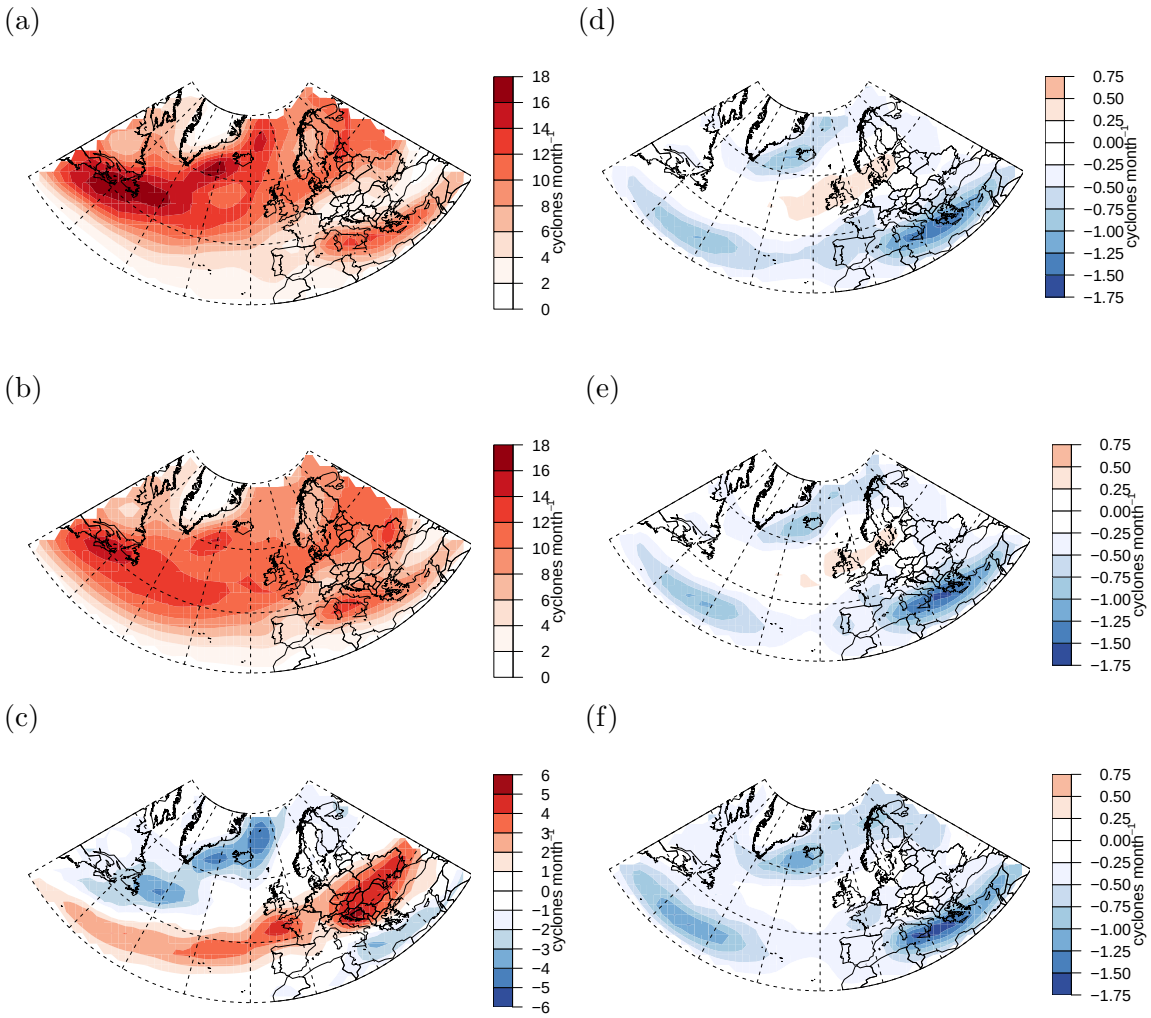

Figure 3.4.: (a) DJF track density in ERA-Interim; (b) CMIP5 expected historical DJF track density estimate from the framework with interactions; (c) difference between CMIP5 and ERA-Interim. Expected climate response estimates from (d) the framework with interactions; (e) the two-way framework; (f) the one-way framework.

Greenland and Iceland compared to the reanalysis, while the more zonal branch towards central Europe is too strong. The result of fitting the three linear regression frameworks to the track density data is shown in the right hand column of Figure 3.4. The usual equally weighted multi-model mean is equivalent to the estimate from the framework with interactions in Figure 3.4d. The estimated climate response from the framework with interactions and the simpler two-way framework are very similar. The similarity between these estimates suggests that there is good agreement between models on the climate response, and the simpler two-way framework may be appropriate. However, the one-way framework estimates a more negative response across most of the study region. It also fails to capture the increase in storm activity over Ireland, the United Kingdom and Denmark indicated by the other two frameworks. This suggests that the estimates from the one-way framework are being biased by large differences between the historical climates of the models and should not be trusted.

3.8.2. Cyclone frequency over London

The differences between the ANOVA frameworks are illustrated by a detailed analysis of a single grid box containing London (51.6N, 1.26E). Large differences are visible in the historical climates simulated by the models in Figure 3.5. The historical mean track density of the models ranges between 7.2 and 14.2 cyclones per month for HadGEM2-ES and BCC-CSM1.1 respectively. Where multiple runs are available, the spread appears comparable between models. This suggests that the assumption of constant variance is a reasonable approximation. The models are almost evenly split between those that simulate a small decrease and those that simulate a small increase in track density in the RCP4.5 scenario. The framework with interactions may be required to allow for this variation if it is greater than might be expected due to internal variability.

The differences between the three linear regression frameworks are visible in the fitted values shown in Figure 3.5. Under the framework with interactions, a different climate response is estimated for each model. The maximum likelihood estimates of the individual model mean climates are the sample means of the runs from each model-scenario pair. Under the two-way framework, all models are assumed to have the same climate response. Models with only one run from a scenario may be poorly fitted due to the effect of internal variability. However, in most cases, the confidence intervals for the expected value in each model-scenario pair still includes these single runs. This suggests that the models may actually agree on the climate response (within the limits of internal variability) and the two-way framework is sufficient to describe the variability present in the ensemble. In the one-way framework, the historical mean climate is assumed to be the same for all models. This fails to capture the large differences visible in the historical climates of the models and is clearly a poor fit to the data.

The maximum likelihood estimate of the ensemble expected climate response β_F and associated 95% confidence interval from the framework with interactions is -0.07 (-0.39,+0.24) cyclones per month. The corresponding estimates from the two-way and one-way frameworks are -0.05 (-0.33,+0.23) and -0.36 (-1.17,0.45) cyclones per month respectively. The similarity between both the point and interval estimates from the framework with interactions and the two-way framework suggests that the models may agree on the climate response and the two-way framework provides a good description of the variation in the ensemble. The dramatic increase in the width of the confidence interval from the one-way framework is due to the large differences between the historical mean climates of the models. Those differences have been absorbed into the estimate of the internal variability, inflating the confidence interval.

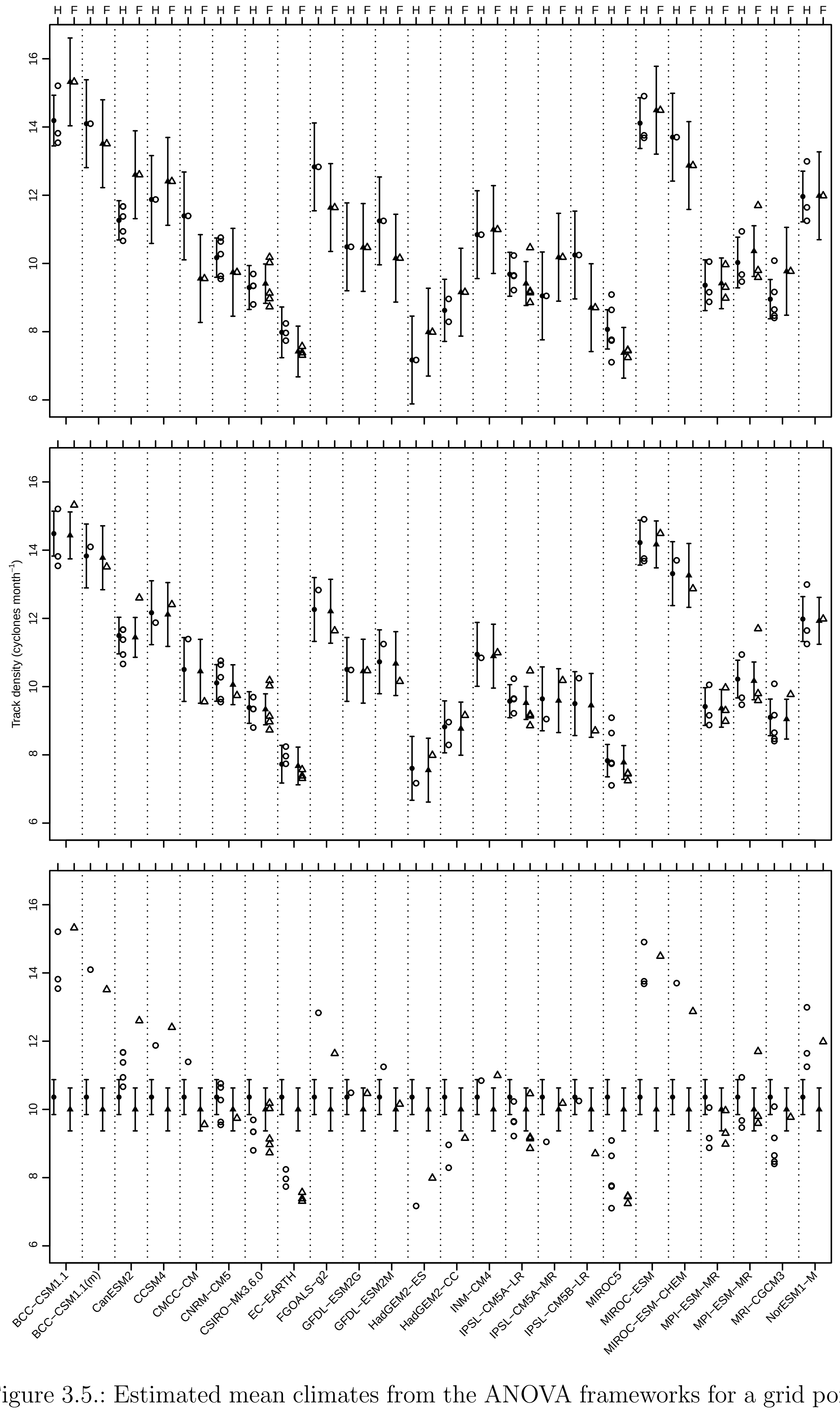

Figure 3.5.: Estimated mean climates from the ANOVA frameworks for a grid point containing London (top) The framework with interactions; (middle) the two-way framework, and (bottom) the one-way framework. Open points represent individual runs from the historical scenario (H, left in each column) and the RCP4.5 (future) scenario (F, right) for each model. Solid points are framework estimates of the mean climate of each model for each scenario. Error bars represent a 95% confidence interval for the mean climate of each model.

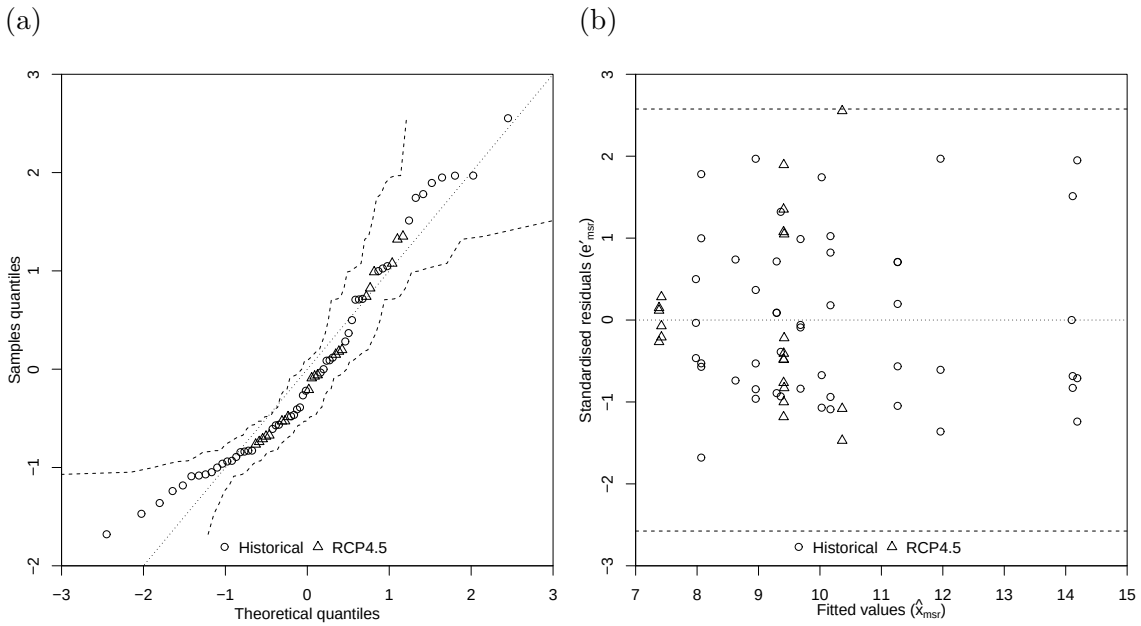

Figure 3.6.: Assumption checking for the framework with interactions, (a) Quantile-quantile plot of the standardised residuals. The dotted line indicates the expected $N(0,1)$ relationship. Dashed lines indicate 95% confidence bounds *on the data* based on a Kolmogorv-Smirnov test. (b) Standardised residuals plotted against fitted values. Dashed lines indicate the 0.5% and 99.5% quantiles of the standard normal distribution.

No systematic patterns and outliers are visible in the plot of standardised residuals against fitted values from the framework with interactions in Figure 3.6b. This suggests that the assumption of constant variance is satisfied at this grid point. The quantile-quantile plot in Figure 3.6a indicates that the residuals are positively skewed compared to a standard normal distribution. The effect is not too pronounced given the small sample size, and the expected normal relationship always lies inside the 95% confidence bounds derived from a Kolmogorov-Smirnov test for normality (Doksum and Sievers, 1976). So the assumption of normality provides a reasonable approximation to the internal variability.

The variance ratio f_γ^2 is calculated as 53%, i.e., differences between the model responses explain approximately half as much variation in the data as the internal variability. The associated estimate of the standardised RMS inter-model spread in the climate response is $\Psi_\gamma = 0.33$ with 95% confidence interval (0.00,1.10). The interval estimate for Ψ_γ includes zero, so there is no significant evidence of model dependence in the climate response at the 5% level, confirmed by the p-value of the F test which is 0.32. Therefore, the simpler two-way framework should provide a good description of the variability in the cyclone track density simulated by the CMIP5 models near London. Rechecking the framework assumptions under the two-way framework reveals no further problems. However, the standardised RMS inter-model spread in the historical climate (Ψ_α) is found to be 2.84 (2.35,3.40). So the models definitely do not agree on historical climate and the one-way framework

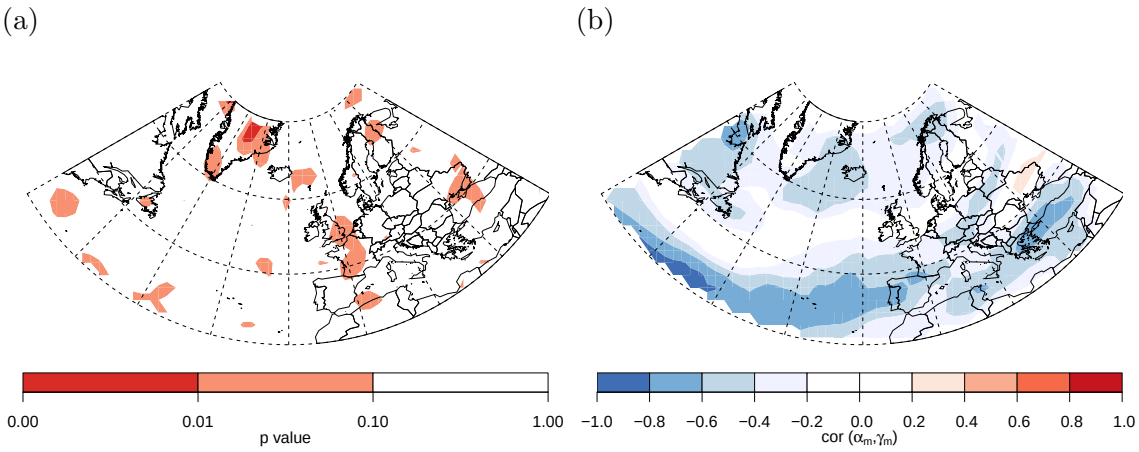

Figure 3.7.: (a) p-values of the Anderson-Darling test for normality in the two-way framework with interactions; (b) Correlation between the estimated climate responses ($\hat{\gamma}_{Fm}$) and historical climates ($\hat{\alpha}_m$) of the models.

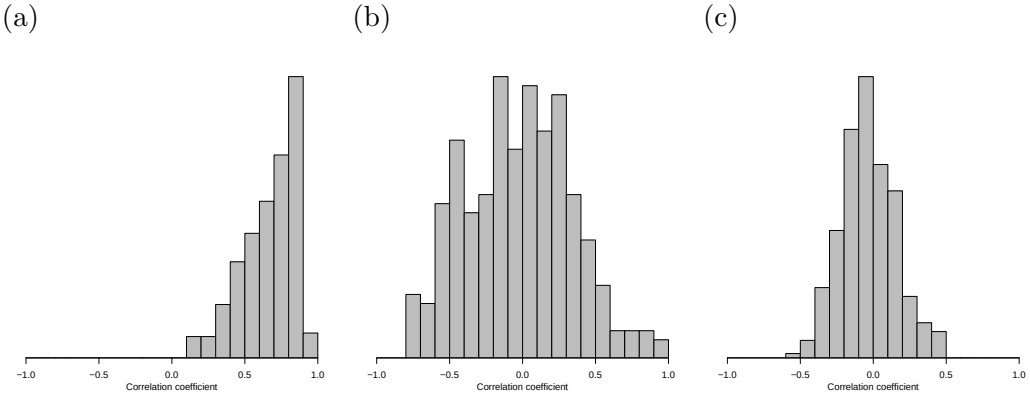

Figure 3.8.: Histograms of correlations between all possible maps of model mean biases in DJF track density in the North Atlantic domain between (a) the historical climates of the CMIP5 models and ERA-Interim; (b) the historical climates of the CMIP5 models and the historical ensemble mean $\hat{\mu}$ from the framework with interactions, i.e., $\hat{\alpha}_m$; (c) the responses of the CMIP5 models and the ensemble mean response $\hat{\beta}_F$ from the framework with interactions, i.e., $\hat{\gamma}_m$.

should not be used.

We can conclude that near London, the CMIP5 models agree reasonably well on the climate response to the RCP4.5 scenario. Under the conditions prescribed by that scenario, there is no significant evidence of a non-zero response in cyclone track density.

3.8.3. The North Atlantic storm track

Comparing the climate response estimates from the three frameworks in Figure 3.4 suggests that the two-way framework may be sufficient to describe the ensemble. However, before the models can be tested for agreement on the climate response, the framework assumptions need to be checked under the two-way framework with

(a) (b)

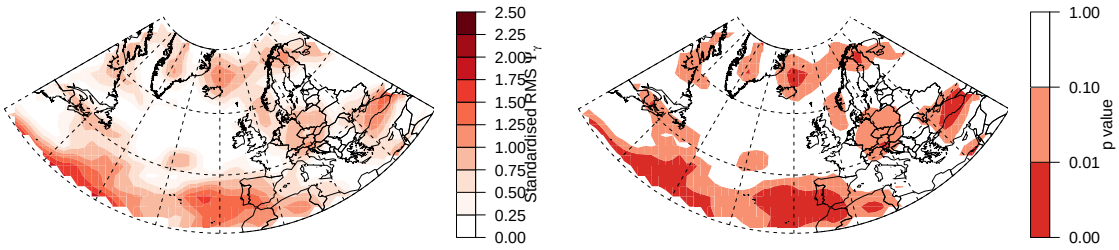

Figure 3.9.: (a) Standardised RMS of the inter-model spread in the climate response (Ψ_γ); (b) p-values of the F tests for model agreement on the climate response.

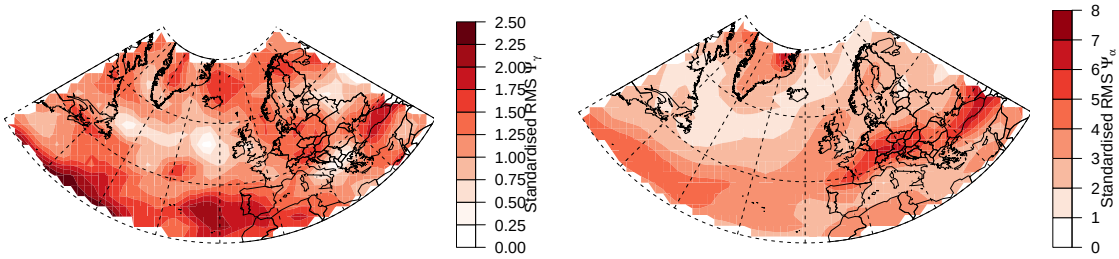

Figure 3.10.: Upper bound of the 95% confidence interval for the standardised RMS inter-model spread in the climate response Ψ_γ.

Figure 3.11.: Standardised RMS of the historical climate Ψ_α.

interactions. Plotting the standardised residuals against the fitted values at a random selection of grid points (not shown) did not reveal any evidence of non-constant variance between models or scenarios. The Anderson-Darling test (Figure 3.7a) suggests that the assumption of normality is acceptable over most of the study region. However, strong correlations between the response ($\hat{\gamma}_{Fm}$) and historical ($\hat{\alpha}_m$) model departures are visible across the ocean basin at 30-40N (Figure 3.7b). If the response departures are dependent on the historical departures, then the models are unlikely to agree on the climate response, otherwise no correlation would be detectable. The potential of this correlation to constrain projections of the future track density response is investigated in Chapter 5. Histograms of the correlations between the historical biases of the models compared to ERA-Interim (Figure 3.8a) suggest the existence of common biases. However, the correlations between the model specific departures α_m and γ_{Fm} (Figures 3.8b and 3.8b) are distributed roughly evenly about zero, suggesting that the models are independently distributed about the ensemble expected climate. In order to identify any outlying runs, the standardised residuals from each run were mapped individually ($N_{\cdot\cdot} = 98$ plots, not shown). No run was found to be outlying at the 1% level at more than 4% of grid points, and those were usually spread over multiple sub-regions, so no further investigation was required.

The framework assumptions appear to be satisfied, so we can examine the evidence of model agreement on the climate response. The estimate of the standardised

RMS Ψ_γ and the p-value of the F test for lack of model agreement are shown in Figure 3.9. The F test indicates significant evidence of model-dependence in the climate response over the sub-tropical North Atlantic between 30-40N. This is consistent with the correlation detected in Figure 3.7b. The standardised RMS model departure Ψ_γ exceeds 1.0 standard deviation over most of this region. Over the ocean between 45-60N the models appear to agree well. However, there is some evidence of model-dependence in the climate response over Central Europe and north of Iceland. The upper bound of the confidence interval on Ψ_γ is shown in Figure 3.10. Where the models agree well in the mid-latitudes, the average standardised model departure is estimated to be less than one standard deviation of internal variability. However, in the sub-tropics, it may exceed two standard deviations.

The t tests on the response departures $\hat{\gamma}_{Fm}$ from the framework with interactions (Figure 3.14) show that CSIRO-Mk3.6.0, FGOALS-g2, MIROC-ESM and MIROC-ESM-CHEM all disagree with the ensemble expected response in the subtropics. This is consistent with Zappa et al. (2013a) who found that the storm tracks in these models were all displaced to the south. As both the historical climate and climate response of these models have been shown to differ from rest of the ensemble, it is possible that they should be excluded from the analysis. The impact of removing models with strongly displaced storm tracks is investigated in Chapter 4.

Since the F test indicates that the models agree well on the climate response outside of the subtropics, we should now check the framework assumptions under the two-way framework before making further inferences. The Anderson-Darling test (not shown) gives no reason to reject the assumption of normality outside of the subtropics. However, in the subtropics between 30N-40N there is strong evidence that the normal assumption is violated. This is unsurprising since the F test indicated significant evidence of model-dependence in the climate response simulated in this region which is not captured by the two-way framework. Checking for outlying runs revealed no cause for concern outside of the same sub-tropical region.

The estimated standardised RMS of the model spread in the historical climate (Ψ_α) is shown in Figure 3.11. The standardised RMS is greater than one standard deviation across the whole study region. The F test for model dependence in the historical climate (not shown) is significant everywhere. In agreement with earlier speculation, this indicates that the one-way framework is not a good description of the variability in the ensemble and should not be used. The models disagree strongly about the track density in the historical scenario over Central Europe, where the standardised RMS model departure approaches eight standard deviations. This is in agreement with Zappa et al. (2013b) who found that the storm tracks of several models were too zonal and extended deep into the European continent. In Appendix A.5, it is shown that $f_\alpha^2 \approx \Psi_\alpha^2$ for large ensembles. So this is also consistent with Sansom

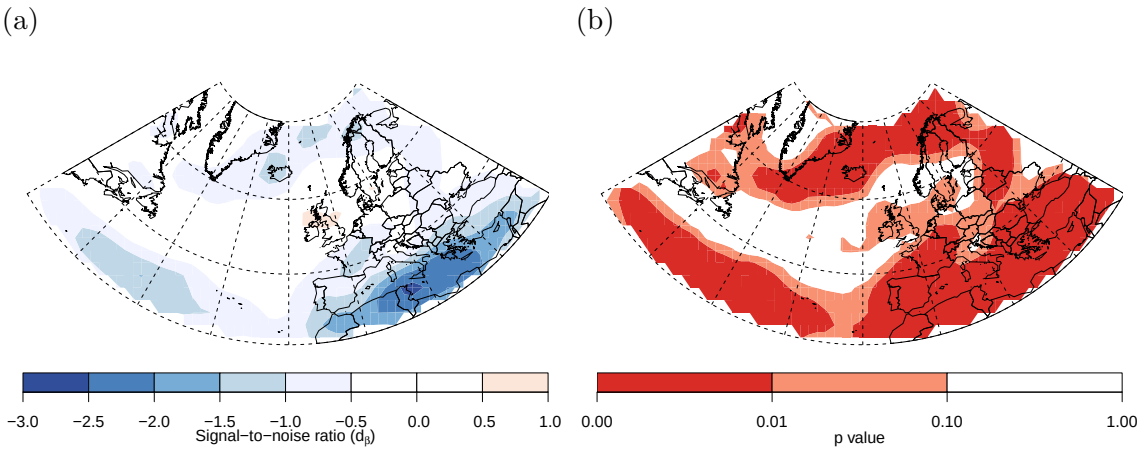

Figure 3.12.: (a) Standardised climate response estimate $\hat{d}_\beta$ from the two-way framework; (b) p-values of the t tests for non-zero climate response from the two-way framework.

et al. (2013) where it was noted that $f_\alpha^2 \approx 70$ over Central Europe.

The F tests suggest that outside of the sub-tropics the models agree well on the climate response and the two-way framework is able to describe the variability in the ensemble. The standardised climate response $\hat{d}_\beta$ and the p-value of the t test for non-zero climate response from the two-way framework are shown in Figure 3.12. The decrease in track density in the subtropical North Atlantic is significantly different from zero at the 1% level. However, there is evidence of poor model agreement and correlation between the climate response and historical climate in this region, so this result should be treated with caution. The standardised climate response is strongest in the Mediterranean basin, reaching almost three standard deviations of internal variability. There is also significant evidence of a small decrease in track density over the North Atlantic between Greenland and Norway.

The differences between the estimates of the expected climate response β_F of the two-way framework and the framework with interactions are shown in Figure 3.13a. The estimates differ most strongly between the Azores and the Iberian Peninsula. The two-way framework assigns a high weight to the responses of CSIRO-Mk3.6.0, EC-EARTH, IPSL-CM5A-LR, MIROC5 and the two models from the MPI (Table 3.1). Five of these six models have positive departures from the ensemble response in this region (not shown) although only CSIRO-Mk3.6.0 is significantly non-zero (Figure 3.14). This explains the weaker estimate of the ensemble expected response under the two-way framework. As expected, where there is no evidence of model dependence in the response, the precision of the estimated climate response in Figure 3.13c is generally greater under the two-way framework than the framework with interactions. Note the much larger decrease in precision when the two-way estimate is used where the models do not agree (the effect is more pronounced in the smaller ensemble of Sansom et al. (2013, Figure 7c)). This is in agreement with

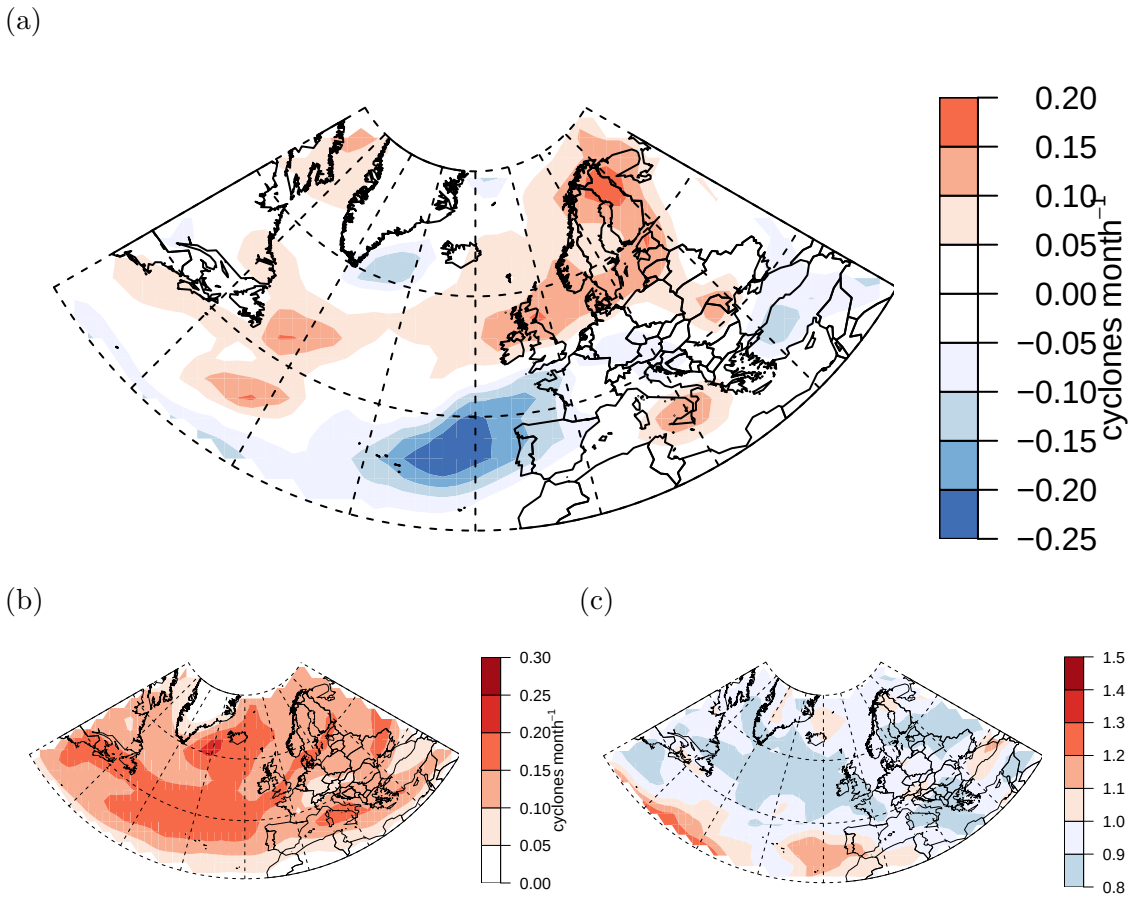

Figure 3.13.: (a) difference between the estimates of the expected climate $\hat{\beta}_F$ from the framework with interactions and the two-way framework; (b) standard error of the estimated expected climate $\hat{\beta}_F$ from the two-way framework; (c) ratio of the standard errors of the expected climate response estimates from the two-way framework and the two-way framework with interactions.

the theoretical arguments of Section 3.3.

Overall, the assumptions of the linear regression frameworks appear to be satisfied outside of the sub-tropical North Atlantic between 30N-40N. In that region, the CMIP5 models suggest a decrease in cyclone frequency of up to one cyclone per month in the winter season. However, this result should be treated with caution since the RMS inter-model spread in the response in this region is of the same order as the simulated response and may be much larger. Elsewhere the models are generally in consensus on the climate response, the inter-model spread usually being small compared to the internal variability. A decrease in cyclone frequency is indicated over the Denmark Strait and Iceland, extending into the Norwegian Sea. There is evidence at the 10% level of a slight increase in activity over Ireland, the United Kingdom and Denmark. The strongest signal is seen in the Mediterranean basin where decreases in cyclone activity of up to 1.75 cyclones per month are simulated. However, Zappa et al. (2013b) found that precipitation in that region may actually increase, so the implications for the water supply in Southern Europe

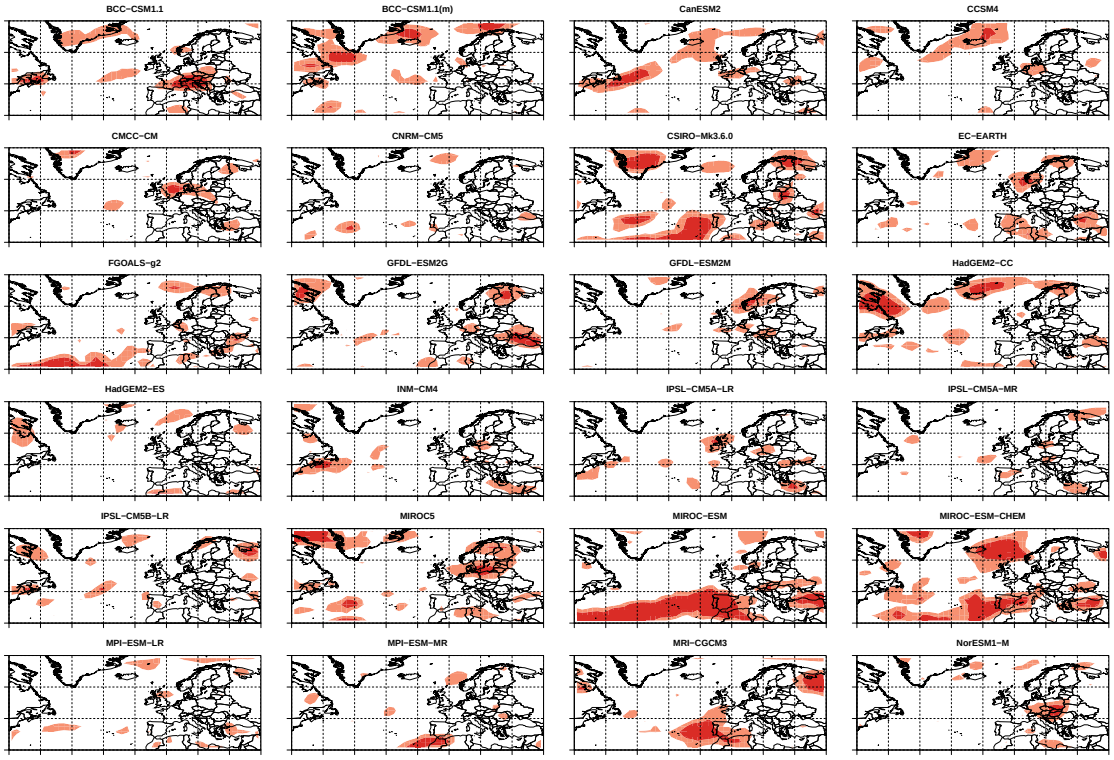

Figure 3.14.: p-values of t tests on individual models for agreement with the ensemble expected response, shading is the same as Figure 3.12b.

may be minimal.

3.9. Discussion

The ANOVA formulation shows that the "one model, one vote" estimate implicitly allows for the possibility that each model will simulate a different response to the same radiative forcing. This makes it difficult to interpret as an estimate of the actual climate response. The principle behind combining information from multiple models is that each model represents a line of evidence for the future state of the climate. If multiple lines of high quality evidence agree, then confidence is increased (Mastrandrea et al., 2010). So if the models all simulate the same climate response, then we should have high confidence in the ensemble expected response as an estimate of the actual climate response. However, the possibility remains that a shared discrepancy may exist between the models and the actual climate (Stephenson et al., 2012). If any discrepancy between the expected clime of the ensemble and the actual climate is not constant over time, then model agreement is not a sufficient criteria for confidence in our projections. The issue of shared discrepancies is discussed in detail in Chapter 6.

The two-way ANOVA framework is estimated under the assumption that the models

all simulate the same climate response, even if they simulate different historical climates. When this assumption is satisfied, the estimate of the expected climate response will usually have greater precision than the "one model, one vote" estimate from the framework with interactions. The estimate of the expected climate response is a weighted average of the responses of the individual models. The model weights depend on the number of runs from both scenarios. Having many runs from only one scenario will not result in a high weight. Modelling centres should therefore aim to provide multiple runs of future scenarios, not only the historical scenario.

The ANOVA frameworks provide a probabilistic description of the variability present in the multi-model ensemble. If the framework assumptions are satisfied, then in principle, a new set of runs could be simulated from the ANOVA frameworks that would be statistically indistinguishable from rerunning the CMIP5 ensemble with the same models but different initial conditions. However, the estimates of the α_m and γ_{Fm} effects are specific to the CMIP5 models and cannot be used to predict the outcome of running a new climate model.

In this chapter, it was shown that F tests based on the ratio of the inter-model spread to the internal variability can be used to test the assumption that the models all simulate the same climate response. A number of effect size estimates have been suggested for F tests (e.g., Cohen, 1973; Olejnik and Algina, 2003; Steiger, 2004). In Sansom et al. (2013), Cohen's f^2 (Cohen, 1988) was used to quantify model agreement. While simple to interpret as a measure of variance explained, it is hard to systematically specify the level of disagreement that we will tolerate on this scale. The standardised RMS of the inter-model spread Ψ has the advantage that we can directly interpret the size of the model differences on the scale of the internal variability.

The most commonly used measure of model agreement is the number of models agreeing on the sign of the ensemble mean response. This measure can also be interpreted in terms of the inter-model spread. The F tests derived here measure agreement on the size of the climate response, as opposed to agreement only on the sign. Agreement on the size of the climate response is a much stricter criteria. However, by separating model differences from internal variability the analysis of the North Atlantic storm track demonstrates that for some variables, agreement between models on the climate response may not be as poor as previously thought. This is in agreement with Tebaldi et al. (2011) who concluded that for some variables, the perceived lack of model agreement was largely due to internal variability masking the signal, i.e., low signal-to-noise ratio in the response. We return to the topic of model agreement in Chapter 4, once we have developed a probabilistic interpretation of the inter-model spread.

Fitting the two-way framework when the models do not agree on the climate response will result in biased estimates and inflated internal variability. F tests such as those outlined here are often used to choose between ANOVA frameworks in order to avoid these problems. However, little attention is usually paid to the power of such tests. If the power is low, then the F tests may fail to detect even quite large differences between the climate models. The approximate confidence intervals described for the standardised RMS of the inter-model spread Ψ allow a far clearer, quantitative assessment of the level of model agreement.

The normal distribution appeared to be reasonable approximation for the internal variability in the 30-year mean cyclone track density data. However, the internal variability of some climate variables will not be normally distributed, even for 30-year averages. If the normal approximation frequently produces physically implausible estimates (e.g., negative frequencies), then a generalized linear model (GLM) could be used in place of a linear regression framework. In a GLM, the internal variability can be modelled by any distribution in the exponential family. Correctly modelling the internal variability is particularly important when events of interest lie in the tails of the distribution. If the distribution is not modelled correctly, then the probabilities assigned to extreme events will be even more strongly affected than events close to the centre of the distribution.

In Sansom et al. (2013) (and in this chapter), the ANOVA frameworks were applied at each grid point independently. Clearly there will be spatial dependence in any climate variable. Therefore a signal that is spatially coherent across a wide area should not necessarily be regarded as more reliable than a more localised one. In climate science, the field significance method (Livezey and Chen, 1983) is sometimes used to account for spatial dependence. However, field significance assumes that the same effect should be expected over the whole study region. Therefore special attention should be paid when defining the region to be tested, e.g., it would not make sense to test the field significance of the change in cyclone track density over a whole hemisphere, since the storm tracks are confined to the mid latitudes. One alternative is the false discovery rate method of Ventura et al. (2004). A better solution would be to build the spatial dependence into the statistical framework. There is a considerable literature on spatial methods in statistics (Cressie, 1993), some of which have already been applied in climate science (e.g., Furrer et al., 2007b). This thesis concentrates on the problem of combining information from multiple models, but does not explicitly consider the spatial structure of the data.

Yip et al. (2011) showed how ANOVA frameworks could be used to quantify the relative contributions of the components of uncertainty in a multi-model ensemble. However, only the internal variability is quantified absolutely. When the contribution of model uncertainty to the climate response is small compared to the internal

variability, then interval estimates based on the two-way framework can be reported for the actual climate response. However, to report interval estimates based on the framework with interactions when the contribution of model uncertainty is not negligible may be misleading. If it is necessary to report such an interval, then it should be accompanied by a statement of limited confidence, as it does not include the contribution from model uncertainty. In Zappa et al. (2013b), we reported results using the framework with interactions for all grid points. When the model uncertainty is small, the estimate of the expected climate response is still unbiased (Appendix A.4) and the internal variability will be well estimated. However, we were careful to point out the regions where the models did not agree and we had limited confidence in our results.

3.10. Conclusion

In this chapter, it was shown that the usual "one model, one vote" estimate of the ensemble expected climate response is equivalent to the maximum likelihood estimate from a two-way ANOVA framework with interactions. Restrictions of this framework yield two alternative estimates, which are more efficient when the model uncertainty is small compared to the internal variability. In contrast to the usual heuristic estimates, the assumptions underlying these estimates are explicit and can be checked using simple graphical techniques. Confidence intervals can be constructed for key parameters and used to quantitatively assess the evidence of a non-zero climate response and the level of agreement between models. The number of models and runs required in order to reliably detect a particular climate response or level of model agreement can be also calculated.

The results contained in this chapter show that it is not always necessary to employ complex Bayesian hierarchical frameworks in order to make rigorous statistical inferences from multi-model ensembles. However, when the model uncertainty is large compared to the internal variability or the climate response depends on the historical climate, a more complex statistical framework is required. In Chapter 4, the ANOVA frameworks are generalised to quantify structural uncertainty as well as internal variability.

4. Quantifying model uncertainty

4.1. Introduction

The ANOVA frameworks described in Chapter 3 can be used to simulate new runs from the models already included in the ensemble. However, unless the models agree on both the historical climate and the climate response, then the ANOVA frameworks cannot be used to simulate runs from new models. If the models do not agree, then the model differences are an additional source of uncertainty, and that uncertainty will not be quantified. Only uncertainty due to internal variability is quantified by the ANOVA frameworks, so the total uncertainty about the climate response may be underestimated. In this chapter, the ANOVA frameworks are extended using random effects to quantify uncertainty due to model differences. Probabilistic representations of the inter-model spread of an ensemble of climate models have been used in a number of previous studies (e.g., Bracegirdle and Stephenson, 2012; Buser et al., 2009; Furrer et al., 2007b; Smith et al., 2009; Tebaldi et al., 2005). However, all of the assumptions underlying this apparently simple construction are rarely stated explicitly. From a frequentist perspective, using random effects to represent model differences would lead us to imagine a notional population of climate models. While it may be possible to imagine such a population, it is difficult to argue that the CMIP5 models represent a random sample from it (Stephenson et al., 2012). Therefore, the ensemble will be reinterpreted from a Bayesian perspective and systematically thinned in order to obtain a set of climate models that can be modelled as a random sample. This leads to a very different picture of model uncertainty in the North Atlantic storm track.

The emphasis in this chapter remains on defining a statistical framework to describe the variation present in the ensemble, and on checking the structure of that framework. In the previous chapter, some simple, but strong, assumptions were identified under which the mean climate response of the ensemble could be interpreted as an estimate of the actual climate. The same assumptions could be applied to the extended framework proposed in this chapter. However, it will be argued that to do so would neglect uncertainty due to shared inadequacies common to all climate models, and that additional assumptions are required in order to relate the ensemble to the

actual climate.

4.2. A hierarchical framework

The ANOVA framework with interactions described in Section 3.3.1 can be rewritten as

$$x_{Hmr} \overset{iid}{\sim} N\left(\mu + \alpha_m \ , \ \sigma^2\right)$$
$$x_{Fmr} \overset{iid}{\sim} N\left(\mu + \alpha_m + \beta_F + \gamma_{Fm} \ , \ \sigma^2\right)$$

since the parameters were constrained such that $\beta_H = 0$ and $\gamma_{Hm} = 0 \ \forall \ m$. Only uncertainty due to internal variability is quantified in this formulation, by the variance σ^2. The uncertainty due to model differences is effectively modelled out by the model departures α_m and γ_{Fm}. In order to quantify model uncertainty directly, a probabilistic description of the model departures is required. Consider the following statistical framework

$$x_{Hmr} \overset{iid}{\sim} N\left(\mu + \alpha_m \ , \ \sigma_H^2\right) \tag{4.1a}$$
$$x_{Fmr} \overset{iid}{\sim} N\left(\mu + \alpha_m + \beta + \gamma_m \ , \ \sigma_F^2\right) \tag{4.1b}$$
$$\alpha_m \overset{iid}{\sim} N\left(0 \ , \ \sigma_\alpha^2\right) \tag{4.1c}$$
$$\gamma_m \overset{iid}{\sim} N\left(0 \ , \ \sigma_\gamma^2\right) \tag{4.1d}$$

Since β now appears only in the expression for future runs x_{Fmr}, the subscript F is dropped for brevity, the same applies to the γ_m terms. The expressions for the historical and future runs x_{Hmr} and x_{Fmr} are otherwise unchanged except for the variance parameters. The assumption of constant internal variability between scenarios is relaxed by including two variance parameters σ_H^2 and σ_F^2, one each for the historical and future scenarios respectively. However, the assumption of constant internal variability between models still applies and the variability in both scenarios is still assumed to be normally distributed. The α_m and γ_m terms still represent the departure of model m from the expected historical climate and climate response respectively, but they are no longer constrained to sum to zero. Instead, a separate statistical model is specified for the α_m terms and for the γ_m terms, forming a second level in what is now a hierarchical framework.

The historical departures α_m are assumed to be independent and identically distributed according to a normal distribution with mean zero and variance σ_α^2. The variance parameter σ_α^2 represents the inter-model spread in the historical climate, and quantifies the model uncertainty in the historical scenario. Since the α_m are assumed to have expectation zero, the expected climates of the models are effec-

tively still centred on μ, but the departures are no longer constrained to sum to zero. Similarly, the variance parameter σ_γ^2 represents the inter-model spread in the climate response, and quantifies the model uncertainty about the response. The response departures γ_m also have expectation zero, so the expected responses of the models are still centred on β.

This formulation represents a considerable simplification compared to the ANOVA framework with interactions in Section 3.3.1. Instead of having to estimate $2M$ parameters, a total of just six parameters now describe the whole ensemble, regardless of how many models are included.

4.3. Assumptions and interpretation

Many of the assumptions underlying the hierarchical model are similar to those of the ANOVA frameworks. The choice of a normal distribution for the internal variability should still be a good approximation for a range of climate variables. The assumption that the internal variability is constant between scenarios has now been relaxed. However, it is still assumed to be constant for all models simulating a particular scenario. Borrowing strength across models in this manner is particularly helpful in the future scenario, where many modelling centres only provide one or two initial condition runs.

Where the hierarchical framework differs from the ANOVA frameworks, is in the assumptions and interpretation surrounding the model specific effects α_m and γ_m. In the ANOVA frameworks, the model specific terms were treated as fixed effects, i.e., a separate parameter was estimated for each model. This implies that the models differ from each other in some systematic way which must be controlled for. However, climate models are inherently similar to one another. They all aim to simulate the behaviour of the same system. They are based on similar sets of equations and numerical codes. The very fact that we wish to combine information from models, implies that we believe that they are similar in some way. Therefore, it seems sensible to describe the model departures by some common parameter.

In the hierarchical framework described by Equation 4.1, the model departures are assumed to be independent samples from identical normal distributions. From the perspective of "classical" frequentist statistics, this implies the existence of a population of models from which those included in the ensemble were sampled. The variances on the model departures represent the spread in the climates simulated by the population of models, much as we might assign a variance to the heights of children of a particular age. However, children grow up and climate models evolve over time. Most of the models in the CMIP5 ensemble are evolutions of models in-

cluded in the CMIP3 ensemble, not independent draws from some super-population of viable models (Stephenson et al., 2012). Therefore, care must be taken when defining the population of models from which the ensemble is sampled. Modelling groups sometimes share developments or even entire model components. The outputs of models that share major components are likely to be more similar than those constructed entirely independently. In statistical terms, the outputs of models with shared components are likely to be correlated. These relationships further complicate the definition of a population of models and a sampling process that might give rise to our ensemble. This makes it difficult to interpret climate models collectively from a frequentist perspective.

From a Bayesian perspective, probability is not restricted to quantifying the frequency with which an event will occur. Rather than defining a population and a sampling process, Bayesian inference relies on judgements about the exchangeability and conditional independence of random quantities, in this case the model outputs. At a basic level, exchangeability simply means that the names of the models are uninformative for their performance (Rougier et al., 2013). This is essentially a restatement of the earlier assumption that the models are all equally valid simulators of the actual climate. The representation theorem (Bernardo and Smith, 2000, pages 177-181) states that a judgement that a set of random quantities are exchangeable can be represented as if they were a random sample from a distribution, conditional on some unknown parameters. So, from a Bayesian perspective, the assumption that the model departures are independent samples from a normal distribution is equivalent to a judgement of exchangeability about the model outputs.

The interpretation of the model outputs as an exchangeable sequence also alters the interpretation of the parameters μ and β. They no longer represent the expected climate and climate response of the models in the current ensemble, but the unknown expectation of an infinite sequence of exchangeable model outputs (effectively a population).

The idea of judging the exchangeability of models is not new to climate scientists. Models are routinely excluded if they exhibit large and inadequately explained biases (e.g., Jun et al., 2008), or once they have been superseded by newer versions with better performance (Knutti et al., 2010a; Reichler and Kim, 2008). It may be necessary to include only a subset of the available ensemble in order to obtain a set of models that we judge to be exchangeable (Rougier et al., 2013). A more rigorous definition of exchangeability requires that we would specify the same marginal distribution for each model departure, as well as the same joint distributions for all possible pairs, triples etc. of departures. Consider all possible pairs of model departures. Having the same joint distribution would imply that the output of each

model has the same covariance (correlation) with the output of every other model. However, the outputs of models from the same modelling group will almost certainly be more similar to each other, than to most other models, i.e., more heavily correlated. Therefore, only one model from each modelling group should be included in the set of models to be analysed. As noted above, models from different modelling groups may also share components. In that case, it may be advisable to include only one model that utilises a particular combination of major components, e.g., a particular pairing of atmosphere and ocean models.

4.4. Fitting the hierarchical framework

The framework described by Equation 4.1 could be classified as a two-way linear mixed model (McCulloch et al., 2008). In Chapter 3, it was assumed that the internal variability was identical in both the historical and future scenarios, i.e., $\sigma_H^2 = \sigma_F^2$. Under that restriction, the hierarchical framework can easily be fitted by maximum likelihood, or restricted maximum likelihood methods, in many statistical software packages, e.g., the lme4 package in the R statistical language. However, ensembles of climate models are more easily interpreted from a Bayesian perspective, and hierarchical frameworks of this kind are easily fitted using Bayesian methods (Gelman et al., 2014).

4.4.1. Prior distributions

The application of Bayes' theorem requires the specification of prior probability distributions for the parameters to be estimated. The prior probabilities express our knowledge or beliefs about the values of the parameters *before* having seen the data. Bayes' theorem then tells us how to update those beliefs given the additional information gained from the data. The quantity that is usually optimised during model fitting in frequentist statistics, the likelihood, is the probability of the data given the parameters. However, the quantity of interest is really the probability of the parameters given the data. Bayes' theorem states that

$$\Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x}\right) \propto \Pr\left(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right)$$

where $\mathbf{x} = (x_{smr} \ \forall \ s, m, r)$ are the model runs, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_M, \gamma_1, \ldots, \gamma_M)$ is the vector of random effects, and $\boldsymbol{\phi} = \left(\mu, \beta, \sigma_H^2, \sigma_F^2, \sigma_\alpha^2, \sigma_\gamma^2\right)$ is the vector of parameters. The probability of interest, $\Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x}\right)$, is referred to as the posterior probability of the parameters. The random effects are estimated jointly with the parameters, but it is the parameters that are of primary interest. The posterior distribution is

proportional to the product of the likelihood $\Pr(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\phi})$ and the prior probability of the parameters and random effects $\Pr(\boldsymbol{\theta}, \boldsymbol{\phi})$. Therefore, flat or vague priors, also called uninformative priors, that give approximately equal probability to all values of the parameters $\boldsymbol{\phi}$ are often chosen, so that prior beliefs have a minimal influence on the posterior probabilities. However, even flat priors are not totally uninformative. A prior specified on a different scale, e.g., $\Pr(1/\phi)$ rather than $\Pr(\phi)$, will result in different inferences for the posterior. Jeffrey's priors are formulated to be invariant to changes of scale, and so have even less influence on the posterior probabilities (Gelman et al., 2014, Section 2.8). Other conditions for uninformative priors are possible, see Bernardo and Smith (2000, Section 5.4) for further examples.

Choosing informative priors can be controversial. Science strives to be objective, and incorporating prior information is often seen as a violation of that principle (Howson and Urbach, 1993). Informative priors have been used in previous studies of multi-model ensembles (e.g., DelSole, 2007; Buser et al., 2009), but the use of flat priors is more common (e.g., Tebaldi et al., 2005; Greene et al., 2006; Smith et al., 2009). The purpose of the hierarchical framework in Equation 4.1 is to provide a probabilistic description of the range of possible model outputs. Therefore flat priors are used throughout in order to minimise the impact of prior belief on our assessment of the behaviour of the models.

Proper priors are specified for all unknown parameters in order to avoid the risk of obtaining an improper posterior distribution. In addition to being proper and flat, the priors are chosen to be conjugate, i.e., the posterior density will belong to the same family of probability distributions as the prior (Bernardo and Smith, 2000, Section 5.2). Conjugate priors have two major advantages over other possible choices. Firstly, they are often easily interpreted as additional data, which makes them a natural choice for specifying prior information, where appropriate. Secondly, it can be shown that the joint posterior distribution of the unknown parameters is determined by the full conditional posterior distributions of the individual parameters (Besag, 1974). Using conjugate priors throughout guarantees that the distribution of each parameter, conditional on all the others, will have a closed form. Therefore, even if no closed form exists for the joint posterior, it can be approximated by sampling from each of the full conditional distributions of the parameters in turn. This process of alternating conditional sampling is known as Gibbs sampling (Bernardo and Smith, 2000, Section 5.5.5).

Diffuse normal priors are specified for the mean parameters

$$\mu \sim N\left(a_\mu\ ,\ b_\mu^{-1}\right) \tag{4.2a}$$

$$\beta \sim N\left(b_\beta\ ,\ b_\beta^{-1}\right) \tag{4.2b}$$

where $a_\mu = a_\beta = 0$ and $b_\mu = b_\beta = 10^{-3}$ so that the mean parameters both have expectation 0 and variance 1000. The prior variance is chosen so that the prior support is much greater than the range of values considered plausible for the parameters. If the parameters were believed to take very large values, e.g., the radius of a cyclone in km, then an even greater prior variance might be required in order to remain broadly uninformative. Conjugate priors are chosen for the computational advantage of being able to sample from the joint posterior. The normal priors combined with the normal likelihood from Equation 4.1 will result in conditional posterior distributions for μ and β that are also normally distributed. This choice reflects a judgement that the mean parameters are expected to be approximately symmetrically distributed with infinite support.

For Bayesian computation, it is convenient to parameterise the normal distribution in terms of the reciprocal of the variance, the precision. The hierarchical framework can be rewritten as:

$$x_{Hmr} \overset{iid}{\sim} N\left(\mu + \alpha_m \ , \ \tau_H^{-1}\right) \tag{4.3a}$$

$$x_{Fmr} \overset{iid}{\sim} N\left(\mu + \alpha_m + \beta + \gamma_m \ , \ \tau_F^{-1}\right) \tag{4.3b}$$

$$\alpha_m \overset{iid}{\sim} N\left(0 \ , \ \tau_\alpha^{-1}\right) \tag{4.3c}$$

$$\gamma_m \overset{iid}{\sim} N\left(0 \ , \ \tau_\gamma^{-1}\right) \tag{4.3d}$$

where

$$\tau_H = \sigma_H^{-2} \ ; \ \tau_F = \sigma_F^{-2} \ ; \ \tau_\alpha = \sigma_\alpha^{-2} \ ; \ \tau_\gamma = \sigma_\gamma^{-2}$$

then vague gamma priors are specified for the precision parameters

$$\tau_H \sim Gamma\left(c_H \ , \ d_H\right) \tag{4.4a}$$

$$\tau_F \sim Gamma\left(c_F \ , \ d_F\right) \tag{4.4b}$$

$$\tau_\alpha \sim Gamma\left(c_\alpha \ , \ d_\alpha\right) \tag{4.4c}$$

$$\tau_\gamma \sim Gamma\left(c_\gamma \ , \ d_\gamma\right) \tag{4.4d}$$

where $c_H = c_F = c_\alpha = c_\gamma = d_H = d_F = d_\alpha = d_\gamma = 10^{-3}$ so that the prior distributions of the precision parameters all have expectation 1 and variance 1000. The large prior variance is chosen for the same reasons given for the mean parameters above. Setting the prior expectation to 1 is a common choice, but it seems appropriate since we do not expect either the model spread or the internal variability to be very small, or very large. Once again, this choice may depend on the scale of the variables being considered.

No closed form exists for the joint distribution of the parameters of the hierarchical framework given the prior distributions specified above. The full conditional pos-

terior distributions of the parameters are derived in Appendix B so that the joint posterior distribution can be approximated by Gibbs sampling.

4.4.2. Initial values

Starting values for each parameter (including the α_m and γ_m) must be specified in order to initialise the Markov chain that will sample from the joint posterior. The maximum likelihood estimates from the ANOVA framework with interactions can be used for the mean parameters μ, β, α_m and γ_m (Appendix A.1). Sample estimates can be used for the precision parameters

$$\tau_H = \frac{R_{H.} - M}{\sum_{m=1}^{M} \sum_{r=1}^{R_{Hm}} \left(\overline{x}_{Hm.} - \alpha_m - \mu\right)^2} \tag{4.5a}$$

$$\tau_F = \frac{R_{F.} - M}{\sum_{m=1}^{M} \sum_{r=1}^{R_{Fm}} \left(\overline{x}_{Fm.} - \gamma_m - \beta - \alpha_m - \mu\right)^2} \tag{4.5b}$$

$$\tau_\alpha = \frac{M-1}{\sum_{m=1}^{M} \alpha_m^2} \tag{4.5c}$$

$$\tau_\gamma = \frac{M-1}{\sum_{m=1}^{M} \gamma_m^2} \tag{4.5d}$$

While the estimates suggested here are obvious choices for the initial values, it is a good idea to try a range of starting values. It is possible that the Markov chain may fail to converge to the proper stationary distribution. If alternative starting values are not sufficient to make the chain converge, then it may be necessary to consider alternative sampling strategies or distributional assumptions.

4.5. Inference in the hierarchical frameworks

The joint posterior distribution of the parameters can be approximated to any required level of precision, simply by obtaining additional samples from the full conditional distributions. The marginal distribution of a particular parameter is approximated by the empirical distribution of all samples of that parameter.

4.5.1. Point estimates

It can be seen from the form of the full conditionals in Appendix B that by choosing vague priors, the influence of the prior information is minimised and the estimates are dominated by the likelihood. Therefore, the modes of the posterior distributions should approximate the maximum likelihood estimates of the parameters. The mode is sometimes called the *maximum a posteriori* (MAP) estimate of a parameter. If

the joint posterior had a known form, then the modes could be found analytically by finding stationary points in the marginal distributions. However, since only the conditional posterior distributions of the parameters are known, the modes would have to be approximated numerically from the samples. The mode is not the only quantity that could be chosen to summarise the value of a parameter. In Bayesian estimation, point estimates are usually made with respect to a loss function based on a measure of the distance between the estimate and the true value (Bernardo and Smith, 2000, p. 257-258). The mode is the quantity that minimises the expected loss according to a zero-one loss function. The mean minimises the expected loss with respect to a quadratic loss function. So the mean of the posterior is the quantity that minimises the mean-squared error of the estimate. If a particular impact of a climate change scenario is of interest, then a specific loss function may be applicable (e.g., the loss associated with not building adequate flood defences). In the absence a specific impact scenario and associated loss function, a symmetric loss function such as the quadratic loss is a sensible choice (Berger, 1985, p. 60-62). Therefore, when reporting point estimates from the hierarchical frameworks, the mean of the samples from the posterior of the parameter in question will be used. If the posterior distribution is symmetric and uni-modal, then the mean and the mode will coincide. The conditional distributions of the mean parameters μ and β are normal, so the means of the posteriors should be close to the maximum likelihood estimates. However, the precision parameters have gamma distributions and so will be positively skewed, as will the posterior distributions of the variance parameters. Therefore, the mean estimates of the variance parameters will tend to exceed equivalent maximum likelihood estimates.

4.5.2. Credible intervals

Credible intervals for any of the parameters may be obtained from the sample quantiles of the simulated conditional posterior distributions. Credible intervals differ in their interpretation from the frequentist concept of confidence intervals. The proper interpretation of a confidence interval is that if a large number of samples were taken, and 95% confidence intervals constructed from each sample for the true value of a parameter, then 95% of those intervals would contain the true value. It does not tell us anything about the probability that any one of those intervals contains the true parameter, it either does, or it does not. On the other hand, a 95% credible interval for β is precisely the interval in which the true parameter is believed to lie with probability 95%. While credible intervals have a far more natural interpretation than confidence intervals, they are not unique for a given posterior distribution. Unless otherwise stated, the $100(1-\alpha)\%$ credible interval for a parameter will be the interval between the $\alpha/2$ and $1-\alpha/2$ quantiles of the conditional posterior distribu-

tion of that parameter. This is an equal tailed credible interval. If the posterior is symmetrically distributed and uni-modal, then the equal tailed interval will coincide with the optimal highest probability density (HPD) credible interval (Bernardo and Smith, 2000, p. 259-262).

In Chapter 3, the use of confidence intervals was encouraged for inference about parameters, rather than relying on simple hypothesis tests The same approach is encouraged for the hierarchical model. Credible intervals for the expected climate response β will be of particular interest. The t test for non-zero climate response in Chapter 3 can be approximated by exploiting the duality between confidence intervals and hypothesis tests (Garthwaite et al., 2002, Section 5.2.3). If zero does not lie within the $100(1-\alpha)\%$ credible interval for β, then we would reject the null hypothesis that $\beta = 0$ at the $100\alpha\%$ level. This does not constitute a true Bayesian hypothesis test (Bernardo and Smith, 2000, p. 262-263), however it is correct to regard a climate response of zero as implausible if it lies outside of the credible interval.

It is now possible to make probability statements about the parameters. For instance, an insurer might be interested in the probability that the expected climate response β will exceed B cyclones per month in London. From a frequentist perspective, such a statement has no meaning since the parameter β is considered to be a fixed quantity. However, from a Bayesian perspective, β is simply another random quantity whose value is uncertain. Therefore, the probability is easily computed from the samples of the posterior distribution of β as

$$\Pr(\beta > B) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{I}\left(\beta^{(n)} > B\right) \tag{4.6}$$

where N is the total number of samples, I is the indicator function and $\beta^{(n)}$ is the nth sample from the posterior distribution of β. Using the credible interval approximation described above, the p-value of the t test for non-zero climate response in Chapter 3 can be approximated by

$$2 \times \min\left(\Pr(\beta > 0)\ ,\ 1 - \Pr(\beta > 0)\right) \tag{4.7}$$

i.e., the probability that the value of β is more extreme than 0.

4.5.3. Model agreement and framework selection

The most commonly used measure of model consensus is the number of models that agree on the sign of the mean response of the ensemble β. This is equivalent to evaluating $\Pr(X_{Rm} > 0)$ (or $\Pr(X_{Rm} < 0)$), where $X_{Rm} = \beta + \gamma_m$ is the expected

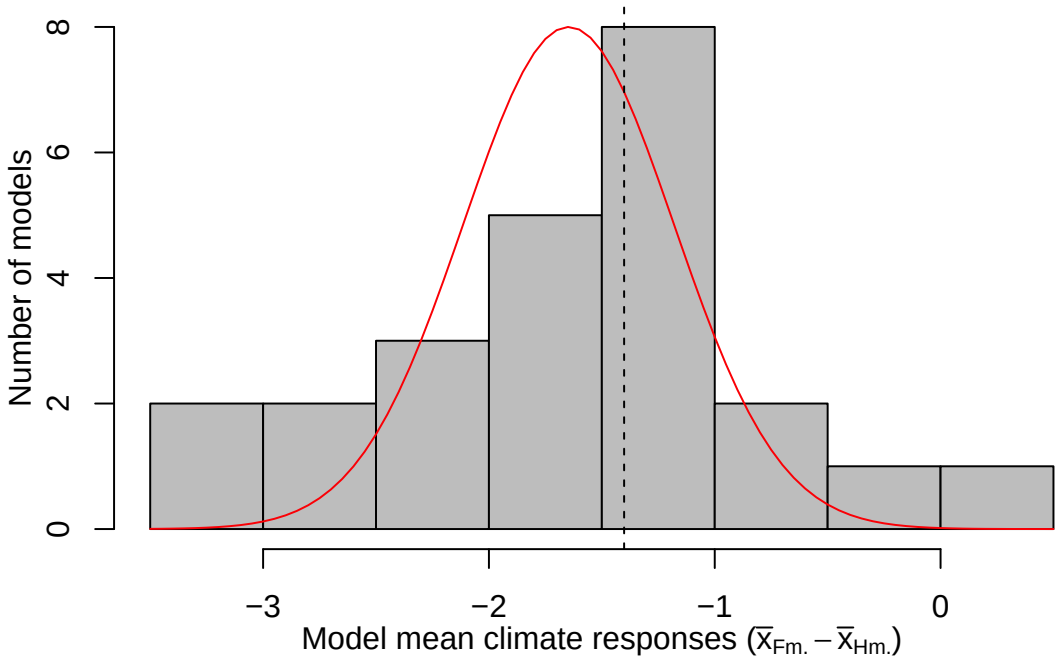

Figure 4.1.: Histogram of model mean cyclone track density responses $\bar{x}_{Fm.} - \bar{x}_{Hm.}$ for a grid box in the Mediterranean (21.4E,36.5N). The red line represents the posterior density of the expected responses of the models ($X_{Rm} = \beta + \gamma_m$), estimated using the hierarchical framework. The black dashed line represents two standard deviations of internal variability away from zero response ($2\sigma_H$).

response of model m, from the empirical probability distribution defined by the model responses (Figure 4.1). This probability can also be evaluated from the samples of the posterior distribution of the parameters β and σ_γ^2.

The IPCC Fifth Assessment Report (Stocker et al., 2013) attempted to separate lack of climate change signal (small change undetectable over internal variability) from lack of model agreement. Regions where the ensemble mean change (β) exceeded two standard deviations of internal variability σ_H *and* 90% of the models agreed on the sign of the response were classed as "large change with high model agreement". Regions where the ensemble mean change (β) was less than one standard deviation of internal variability were classed as "small signal or low agreement of models". Due to the fact that only one run was used from each model, model differences were compounded by internal variability and lack of signal still could not really be distinguished from lack of model agreement. The framework derived here overcomes this by separating model uncertainty from internal variability. Therefore, model agreement on either the sign *or* the size (as suggested in Chapter 3) of the climate response can be evaluated independent of the internal variability.

In the Fifth Assessment Report (Stocker et al., 2013), a large change was defined as one that exceeds two standard deviations of internal variability, i.e., one that is detectable (statistically significant) when compared to the recent climate. Tebaldi

et al. (2011) argue that "evaluating model agreement is only meaningful if the models are producing significant changes". This is equivalent to quantifying agreement on the climate response based on $\Pr(X_{Rm} > 2\sigma_H)$, which is also easily evaluated from the samples of the posterior distribution of the parameters (Figure 4.1). This definition more closely resembles the measure of agreement derived in Chapter 3, since it considers the size of the model differences relative to the internal variability. However, it still depends on the expected size of the response in addition to the inter-model spread. The F tests derived in Chapter 3 provide a pure measure of agreement, independent of the size of the response.

In Chapter 3, the F tests were used to test for model consensus and so choose which ANOVA framework to make inferences from. It would be far simpler to have a single framework which could be used regardless of how well the models agree on key parameters. The hierarchical framework provides that simplicity. The model uncertainty is represented by the variance parameters σ_γ^2 and σ_α^2. So, no matter how large or small that uncertainty is, it is quantified. Most climate modellers would accept that models will never agree completely. Different choices of numerical methods, underlying equation sets and process parameterisations will inevitably lead to differences in the simulated outputs. Therefore, it seems more natural to simply allow for the differences, rather than to complicate the process of analysing the ensemble by asking whether or not the models agree *exactly*. Credible intervals may be obtained for the variances of the model departures σ_α^2 and σ_γ^2. These are interpreted in a similar manner to the confidence intervals for the standardised RMS of the inter-model spread in Chapter 3. If the mean of the posterior of the variance is small but the credible interval includes very large values, then the inter-model spread is not well determined and the model consensus may actually be poor.

4.5.4. Prediction

The posterior predictive distribution of new runs $\tilde{\mathbf{x}}$ given the runs already in the ensemble is

$$\begin{aligned}
\Pr(\tilde{\mathbf{x}} \mid \mathbf{x}) &= \iint \Pr(\tilde{\mathbf{x}}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x})\, d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&= \iint \Pr(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) \Pr(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x})\, d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&= \iint \Pr(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \Pr(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x})\, d\boldsymbol{\theta} d\boldsymbol{\phi}
\end{aligned}$$

since the outcome of new runs $\tilde{\mathbf{x}}$ is independent of the existing runs $\mathbf{x}$ given the random effects $\boldsymbol{\theta}$ and the parameters $\boldsymbol{\phi}$. After sampling from the full conditional distributions of the parameters using the Gibbs sampler, N samples from the pos-

terior distribution $\Pr(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x})$ are effectively available, and $\Pr(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$ is given by Equations 4.1a & 4.1b, so prediction from the hierarchical model is simple, as described below.

Predicting runs from existing models

From Equation 4.1, the posterior predictive distribution for a new run of existing model m from the historical scenario can be approximated by drawing one sample from

$$\tilde{x}^{(n)}_{Hmr} \mid \mathbf{x} \sim N\left(\mu^{(n)} + \alpha^{(n)}_m \ , \ \sigma^{2\,(n)}_H\right) \tag{4.8}$$

for each of the $n = 1, \ldots, N$ samples of the joint posterior of the parameters. Similarly, the posterior predictive distribution for a new run of existing model m from the future scenario can be approximated by drawing one sample from

$$\tilde{x}^{(n)}_{Fmr} \mid \mathbf{x} \sim N\left(\mu^{(n)} + \alpha^{(n)}_m + \beta^{(n)} + \gamma^{(n)}_m \ , \ \sigma^{2\,(n)}_F\right) \tag{4.9}$$

for each of the N samples of the joint posterior of the parameters. Finally, the posterior predictive distribution of the difference between a future and a historical run from existing model m, i.e., the climate response of model m, can be approximated by differencing the samples from Equations 4.9 and 4.8.

Predicting runs from new models

The ANOVA frameworks in Chapter 3 are able to predict the outcomes of new runs from models already in the ensemble, but not the outcomes of runs from a new model. Using the hierarchical model, it is also simple to predict the output of a new model that is judged to be exchangeable with the models already in the ensemble. For a new model j, the random effects representing the model departures $\tilde{\boldsymbol{\theta}}_j = (\tilde{\alpha}_j, \tilde{\gamma}_j)$ are unknown. The posterior predictive distribution for the departure of a new model from the expected climate is

$$\Pr\left(\tilde{\boldsymbol{\theta}}_j \mid \mathbf{x}\right) = \int \Pr\left(\tilde{\boldsymbol{\theta}}_j \mid \boldsymbol{\phi}\right) \Pr(\boldsymbol{\phi} \mid \mathbf{x})\, d\boldsymbol{\phi}$$

since the departure $\tilde{\boldsymbol{\theta}}_j$ of a new model from the expected climate is independent of the existing runs given the parameters $\boldsymbol{\phi}$. N samples from the joint posterior distribution of the parameters $\Pr(\boldsymbol{\phi} \mid \mathbf{x})$ are already available and $\Pr\left(\tilde{\boldsymbol{\theta}}_j \mid \boldsymbol{\phi}\right)$ is given by Equations 4.1c & 4.1d. So the posterior predictive distribution for a new run of the historical scenario by *unknown* model j, can be approximated by first

sampling a new historical departure $\tilde{\alpha}_j$ from

$$\tilde{\alpha}_j^{(n)} \mid \mathbf{x} \sim N\left(0\ ,\ \sigma_\alpha^{2\,(n)}\right) \tag{4.10}$$

for each of the $n = 1, \ldots, N$ samples from the posterior distribution of the parameters. Then sample a new run from

$$\tilde{x}_{Hjr}^{(n)} \mid \mathbf{x} \sim N\left(\mu^{(n)} + \tilde{\alpha}_j^{(n)}\ ,\ \sigma_H^{2\,(n)}\right) \tag{4.11}$$

for each of the N samples of the posterior distribution of the parameters. Similarly, the posterior predictive distribution for a new run of the future scenario by *unknown* model j can be approximated by first sampling a new historical departure $\tilde{\alpha}_j$ from Equation 4.10, and then sampling a new response departure $\tilde{\gamma}_j$ from

$$\tilde{\gamma}_j^{(n)} \mid \mathbf{x} \sim N\left(0\ ,\ \sigma_\gamma^{2\,(n)}\right) \tag{4.12}$$

for each of the $n = 1, \ldots, N$ samples from the posterior distribution of the parameters. Then, sample a new run from

$$\tilde{x}_{Fjr}^{(n)} \mid \mathbf{x} \sim N\left(\mu^{(n)} + \tilde{\alpha}_j^{(n)} + \beta^{(n)} + \tilde{\gamma}_j^{(n)}\ ,\ \sigma_F^{2\,(n)}\right) \tag{4.13}$$

for each of the N samples from the posterior distribution of the parameters. Finally, the posterior predictive distribution of the difference between two runs of *unknown* model j, i.e., the climate response of model j, can be sampled by differencing samples from Equation 4.13 and Equation 4.11 for the same historical departure $\tilde{\alpha}_j^{(n)}$, or by sampling directly from

$$\widetilde{x_{Fjr} - x_{Hjr'}}^{(n)} \mid \mathbf{x} \sim N\left(\beta^{(n)}\ ,\ \sigma_\gamma^{2\,(n)} + \sigma_H^{2\,(n)} + \sigma_F^{2\,(n)}\right) \tag{4.14}$$

Predicting the actual climate

Under the "truth plus error" paradigm, the expected response of the actual climate would be assumed to coincide with the expected response of the models β. Intuition tells us that if the models do not all simulate the same response, then our uncertainty about the actual climate response should reflect the spread of the responses simulated by the models. Equation 4.14 clearly satisfies that intuition for a new model. However, unless the models agree *exactly* the credible interval for β will always be narrower than the inter-model spread in the responses, since it is the credible interval for the expectation of the model responses (effectively a sample mean).

An alternative to the "truth plus error" approach is to assume that the models

are exchangeable with the actual climate (Annan and Hargreaves, 2010). Such an assumption would imply that the actual climate differs structurally from the models in the same way that the models differ structurally from each other. In that case, it would appear straightforward to include the observed historical climate in the fitting procedure as though it were a run from another model. The posterior predictive distribution of the actual climate response could then be derived, similar to predicting the response of a new model in Equation 4.14.

However, in Chapter 2 it was argued that the actual climate should not be viewed as simply another model. Climate models are discretised approximations based on differential equations, and are fundamentally different from the system they are trying to represent. In addition, while treating the observed climate as a run from a model would account for structural uncertainty and initial condition uncertainty from the models, and sampling uncertainty about the actual climate due to natural variability, it would ignore the uncertainty due to measurement error associated with the observations themselves. We return to the problem of relating the ensemble to the actual climate in Chapter 6. For now we concentrate on ensuring that the hierarchical framework provides a good description of the behaviour of the climate models.

4.6. Framework checking

4.6.1. Convergence

One of the issues with Gibbs sampling is that many iterations can be required before the Markov chain converges to the stationary distribution of the parameters. A simple check for convergence is to examine the time series of samples for each parameter and look for the approximate point at which each distribution stabilises. This is known as the burn-in period. At each step in the chain, the distribution of the next sample of each parameter is conditional on the current values of all the other parameters. Therefore, all samples prior to stabilisation of the parameter with the longest burn-in period should be discarded. Gelman et al. (2014, Section 11.4) suggest running several chains simultaneously, each starting from different initial conditions, and monitoring convergence based on within and between sequence variances for each parameter. This approach may be useful for the automation of fitting to several different datasets, or if it is not clear that the Markov chain has converged. However, it may not be practical to check for convergence at each location when fitting to many grid points. In that case graphical checks at a random selection of grid points should be sufficient to suggest a suitable burn-in period.

4.6.2. Autocorrelation

The other main issue with iterative sampling procedures is that because the distributions are conditional on the current values of all the other parameters, correlations with subsequent samples may be large and persistent. Autocorrelation is not necessarily a problem. All the samples after the burn-in period are still valid samples from the joint distribution of the parameters. However, in the presence of autocorrelation, the total information about the the parameters will be less than is indicated by the number of samples, i.e., the effective number of samples is less than the actual number. The extent of the autocorrelation can be assessed by computing and plotting the autocorrelation function over a large number of lags for each parameter. Ideally, the autocorrelation should drop rapidly with increasing lag and then remain close to zero. One option is to thin the samples by only keeping every kth sample where k is the lag at which the autocorrelations of all the parameters approach zero. By running several chains simultaneously, the within and between chain variances can be also be used to estimate the effective number of samples (Gelman et al., 2014, Section 11.5). In that case, sampling may be halted after the required effective number of samples have been obtained. As noted for convergence, it may not be practical to check the autocorrelation at each location when fitting to a large number of grid points. Graphical checks at a random selection of grid points should be sufficient to suggest a suitable thinning strategy.

4.6.3. Cross-validation

Validating the assumptions involved in hierarchical frameworks such as Equation 4.1 is difficult due to the multi-layered structure. The assumptions about the model departures α_m and γ_m are particularly problematic since they are latent variables and cannot be observed directly. The expected historical climate, future climate and climate response of model m are

$$\begin{aligned} \mathrm{E}\left(x_{Hmr}\right) &= \mu + \alpha_m \\ \mathrm{E}\left(x_{Fmr}\right) &= \mu + \alpha_m + \beta + \gamma_m \\ \mathrm{E}\left(x_{Fmr} - x_{Hmr}\right) &= \beta + \gamma_m \end{aligned}$$

Natural estimates of the expected values are the means of the runs made by model m under each scenario, i.e., $\overline{x}_{Hm.}$, $\overline{x}_{Fm.}$, and their difference $\overline{x}_{Fm.} - \overline{x}_{Hm.}$. Suppose there are N_{Hj} runs of the historical scenario and N_{Fj} runs of the future scenario by a new model j, which is judged to be exchangeable with the M models already in the ensemble. The posterior predictive distribution for the mean of the historical runs from the new model can be approximated by taking the mean of N_{Hj} samples from

Equation 4.11 for each of the N samples from the joint posterior of the parameters. Alternatively, it can be sampled directly from

$$\widetilde{\overline{x}_{Hj.}}^{(n)} \mid \mathbf{x} \sim N\left(\mu^{(n)} \ , \ \sigma_{\alpha}^{2\,(n)} + \frac{\sigma_H^{2\,(n)}}{R_{Hj}}\right) \tag{4.15}$$

Similarly, the posterior predictive distribution for the mean of the future runs from the new model can be approximated by taking the mean of N_{Fj} samples from Equation 4.13 for each of the N samples from the joint posterior of the parameters. It can also be sampled directly from

$$\widetilde{\overline{x}_{Fj.}}^{(n)} \mid \mathbf{x} \sim N\left(\mu^{(n)} + \beta^{(n)} \ , \ \sigma_{\alpha}^{2\,(n)} + \sigma_{\gamma}^{2\,(n)} + \frac{\sigma_F^{2\,(n)}}{R_{Fj}}\right) \tag{4.16}$$

However, the main quantity of interest is the mean climate response of the new model. Its posterior predictive distribution can approximated by sampling from

$$\widetilde{\overline{x}_{Fj.} - \overline{x}_{Hj.}}^{(n)} \mid \mathbf{x} \sim N\left(\beta^{(n)} \ , \ \sigma_{\gamma}^{2\,(n)} + \frac{\sigma_H^{2\,(n)}}{R_{Hj}} + \frac{\sigma_F^{2\,(n)}}{R_{Fj}}\right) \tag{4.17}$$

This suggests a cross validation approach to checking the distributional assumptions about the model departures (the procedure described here is equivalent to that of Smith et al. (2009), see Appendix B.2)

1. For each $j \in 1, \ldots, M$ refit the hierarchical framework, leaving out the runs from model j.

2. For each of the N samples from the new posterior distribution of the parameters without model j, draw one sample from the posterior predictive distribution of the sample mean climate response of model j (Equation 4.17).

3. Calculate the proportion of samples from the posterior predictive distribution that are greater than or equal to the value of the sample mean climate response of model j

$$p_j = \frac{1}{N}\sum_{n=1}^{N} \mathrm{I}\left(\widetilde{\overline{x}_{Fj.} - \overline{x}_{Hj.}}^{(n)} > \overline{x}_{Fj.} - \overline{x}_{Hj.}\right) \tag{4.18}$$

4. Repeat steps 2 & 3 for the mean historical climate of model j using Equation 4.15, and the mean future climate using Equation 4.16, if required.

The quantity calculated in Step 3 is an estimate of the posterior predictive p-value for the mean climate response of model j. Similar to the check for outlying runs in Chapter 3, models with $p_j \leq 0.005$ or $0.995 \leq p_j$ should be examined carefully for inconsistent behaviour. If several models have $p_j \leq 0.005$ or $0.995 \leq p_j$ then the

distributional assumptions about the model departures α_m and γ_m may need to be reassessed.

The cross validation procedure is designed to test the assumptions about the model departures, but it is also influenced by uncertainty due to internal variability. If the internal variability is comparable to the inter-model spread, and the models only simulate one or two runs of each scenario, then the posterior predictive distributions will be influenced as much by internal variability as by the model spread. Therefore, it may be the assumptions about the internal variability that are at fault, not about the model departures.

In principle, similar tests could also be used to test the assumptions about the internal variability by leaving out one run at a time, and then predicting its outcome using the predictive distributions in Section 4.5.4. However, if many grid points must be tested, then the computational cost of such a scheme would be prohibitive for an ensemble of even moderate size. Fitting the ANOVA framework with interactions and performing the residual checks or Anderson-Darling test described in Chapter 3 should be adequate to check the assumptions about the internal variability.

4.7. Application to the North Atlantic storm track

In this section, the hierarchical framework and cross-validation methodology are applied to the North Atlantic storm track data from Chapter 3. The hierarchical framework was initially implemented entirely in the R statistical language. The implementation itself is simple thanks to the extensive random number generation facilities available. However, interpreted languages such as R are not able to handle iterative procedures efficiently. Therefore the framework was reimplemented in Fortran 90 using the random number generators collected by Chandler (2003). R is able to call Fortran subroutines directly once compiled into an appropriate shared object. So the full data handling capabilities of the R language can be leveraged for setting up and analysing the simulations while simultaneously exploiting the efficiency of Fortran for scientific programming.

The updates for the individual parameters all have similar complexity. Therefore the runtime of the Gibbs sampler is approximately proportional to

$$(N \times T + B) \times (P + 2 \times M) \tag{4.19}$$

where N is the number of samples required after thinning by retaining only every Tth sample, B is the burn-in period, $P = 6$ is the number of parameters, and M is

the number of models. Running on a 3.06 GHz desktop PC dating from 2009, the Fortran version takes around 8.0 seconds to generate $N = 10^6$ from the posterior distribution of the parameters samples, with no thinning ($T = 1$), no burn-in ($B = 0$) and $M = 24$ models. This compares to 162 seconds for the same number of samples when implemented natively in R.

4.7.1. Cyclone frequency over London

The checks on the sampling process described in Section 4.6 were carried out at a selection of grid points across the the study region in order to determine appropriate thinning and burn-in strategies. The checking procedures are illustrated here for the grid box containing London (51.6N,1.26E). In testing, the sampler was run for $N = 10^6$ samples with no thinning ($T = 1$). The time series in Figure 4.2 show that the chains for β, τ_H, τ_F and τ_α are all stable and well mixed with little or no apparent burn-in period. The chain of samples of μ is also stable but appears to be mixing more slowly. This is consistent with the autocorrelations in Figure 4.3 where the other variables exhibit rapidly decaying autocorrelations but μ may exhibit significant autocorrelation over 200 samples or more.

More concerning is the time series of samples of τ_γ. The mode of the distribution lies between 4-5, but the chain makes frequent and prolonged excursions into regions of very high precision ($\tau_\gamma \gg 1000$) leading to a highly skewed distribution. The reasons for these excursions are easily understood by considering the full conditional distributions of τ_γ and γ_m. If the models are in good agreement on the response so that $\sum_{m=1}^{M} \gamma_m^2 \ll M$, then both the expectation and variance of τ_γ will be large (Equation B.11 of Appendix B). When τ_γ is large, then both the expectation and variance of all the γ_m will be approximately zero (Equation B.7). In that case, the sum of squares in Equation B.11 will remain small, making it hard for the chain to jump back to small values of τ_γ. But what does this mean? A large value of τ_γ implies that the models agree almost exactly on the climate response. If the model response departures γ_m are small compared to the internal variability, then unless a large number of runs are available, the model differences will not be distinguishable from the internal variability. In that case, $\gamma_m \approx 0 \; \forall \; m$ is an admissible solution and these long excursions into a different area of parameter space are the result.

For some grid points, the distribution of τ_γ actually becomes bi-modal. The sampler is effectively exploring two different frameworks, one where the models agree and one where they do not. Such bi-modal behaviour creates obvious difficulties for inference, e.g., the mean of the distribution will not correspond to either case. Putting aside this objection, is such behaviour realistic? In Chapter 3, the situation where the models agreed was desirable since it removed one source of variability which was

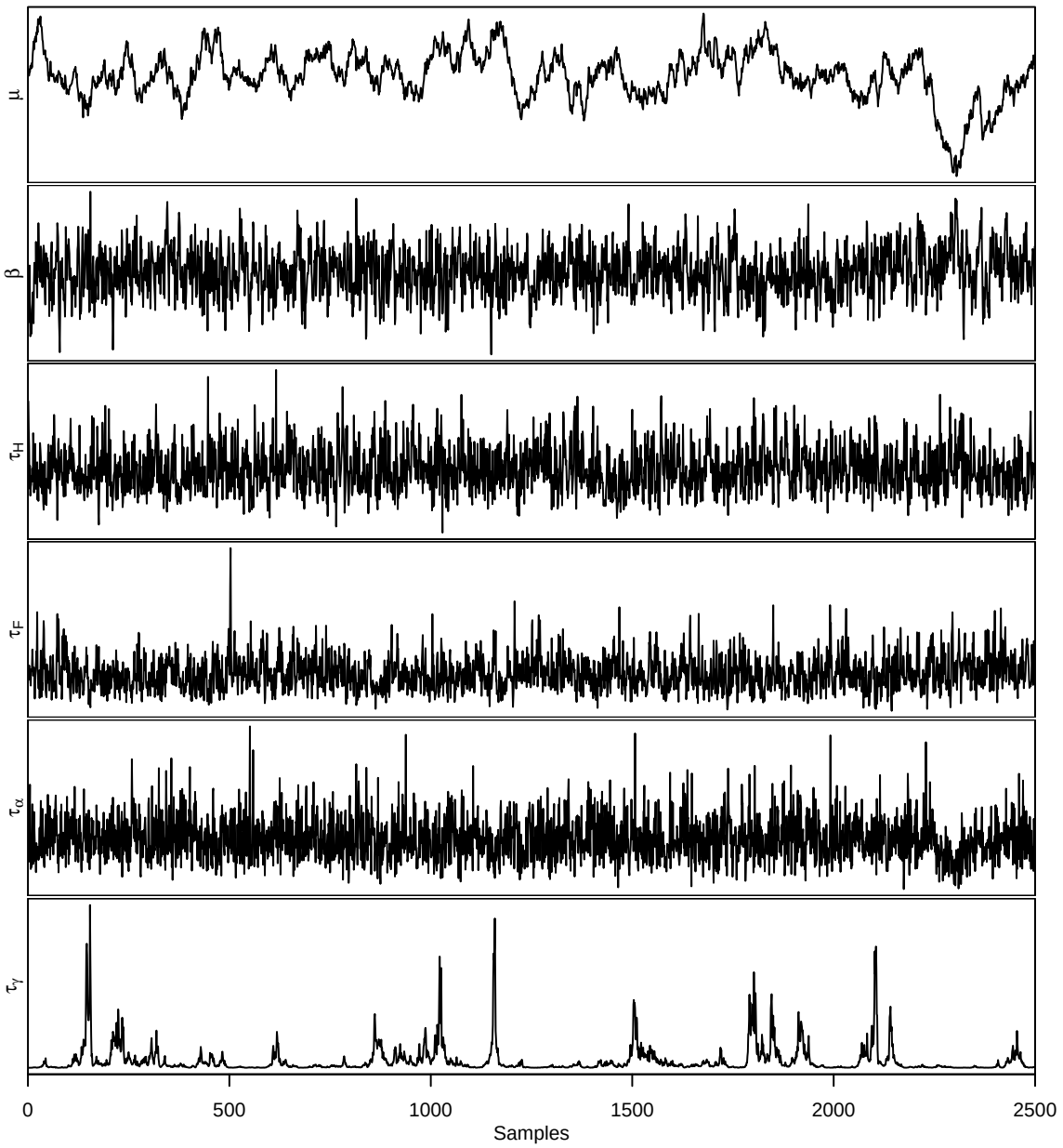

Figure 4.2.: Time series of the first 2500 samples from the joint posterior distribution of the parameters for the grid box containing London.

otherwise unaccounted for. However, given the variety of ways that climate models differ from one another, it seems optimistic to expect that they should ever agree completely. It is easy to incorporate this belief into the hierarchical framework via the prior distribution for τ_γ. It can be seen from Equation B.11 that the prior rate parameter d_γ is effectively a lower bound on the sum of squared model response departures. Suppose that at best, we believe the model responses will have a range of 0.5 cyclones per month. Since the model departures are assumed to arise from a normal distribution, 99.7% of the mass of the distribution lies within three standard deviations of the mean. Therefore the expectation of the sum of squared departures is approximately $M \times (0.5/6)^2 = 0.167$ for $M = 24$, implying $d_\gamma = 0.083 \approx 10^{-1}$.

Further testing showed that $d_\gamma = 10^{-1}$ is sufficient to prevent the sampler from

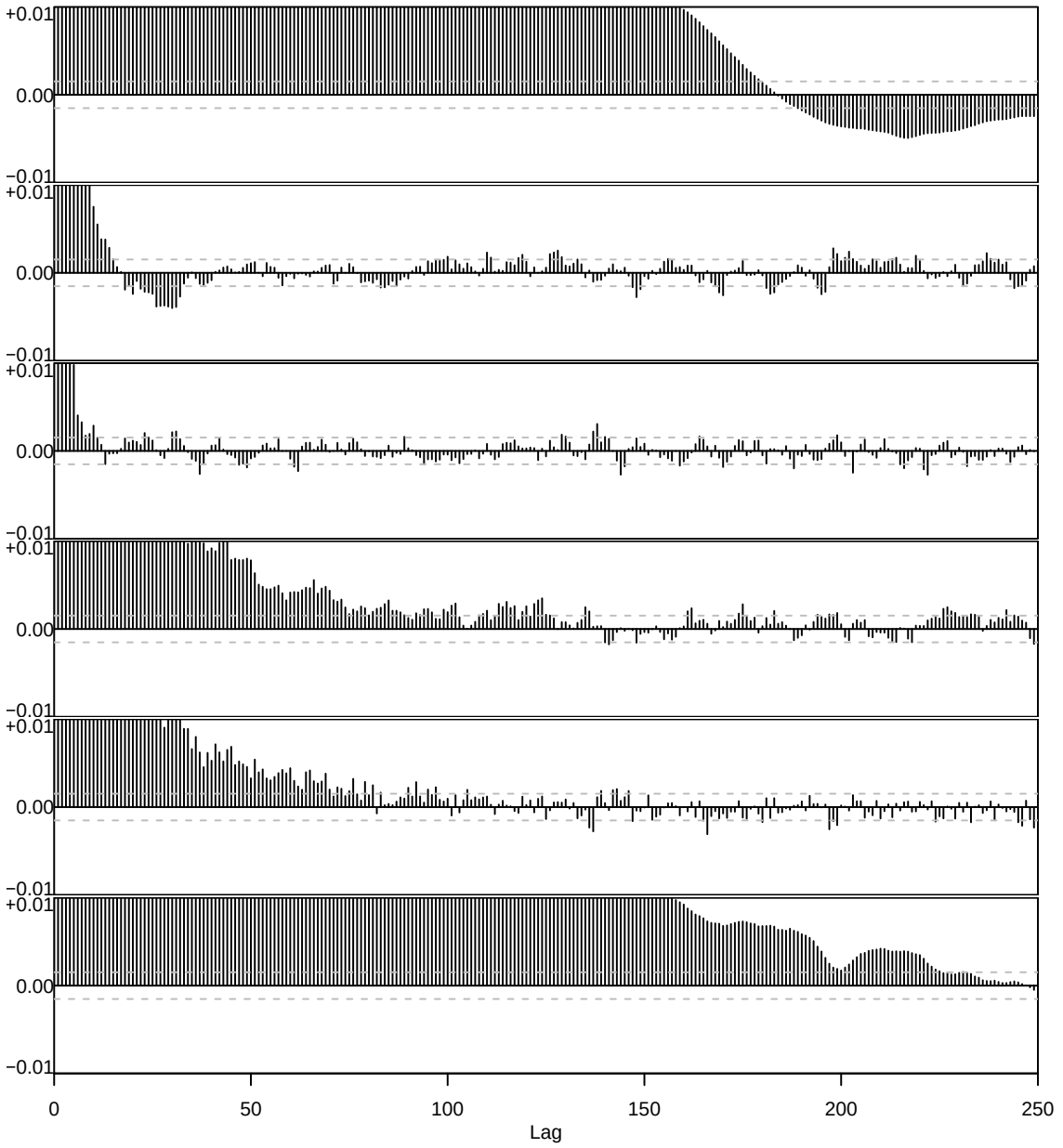

Figure 4.3.: Autocorrelations of samples of, from top to bottom, μ, β, τ_H, τ_F, τ_α, τ_γ. The horizontal dashed lines are 90% confidence intervals for the expected autocorrelation based on a white noise process.

stalling in regions of very high precision. The effect of the informative prior is illustrated in Figure 4.4. If the sum of squared model departures is greater than 3.7 ($\sqrt{\sigma_\gamma^2} \approx 0.39$ cyclones per month for $M = 24$), then the conditional mean and variance of τ_γ lie within 5% and 10% of their respective values under the uninformative prior. So if the models disagree more than slightly, then the effect of the informative prior is negligible. The overall effect of the informative prior is conservative, it tends to underestimate the precision τ_γ and hence overestimate the inter-model spread σ_γ^2.

Figure 4.5 compares the posterior densities of the parameters for using the informative and uninformative priors for τ_γ. The mode of τ_γ remains almost unchanged but the skewness is reduced by the restriction on the sum of squared departures. The

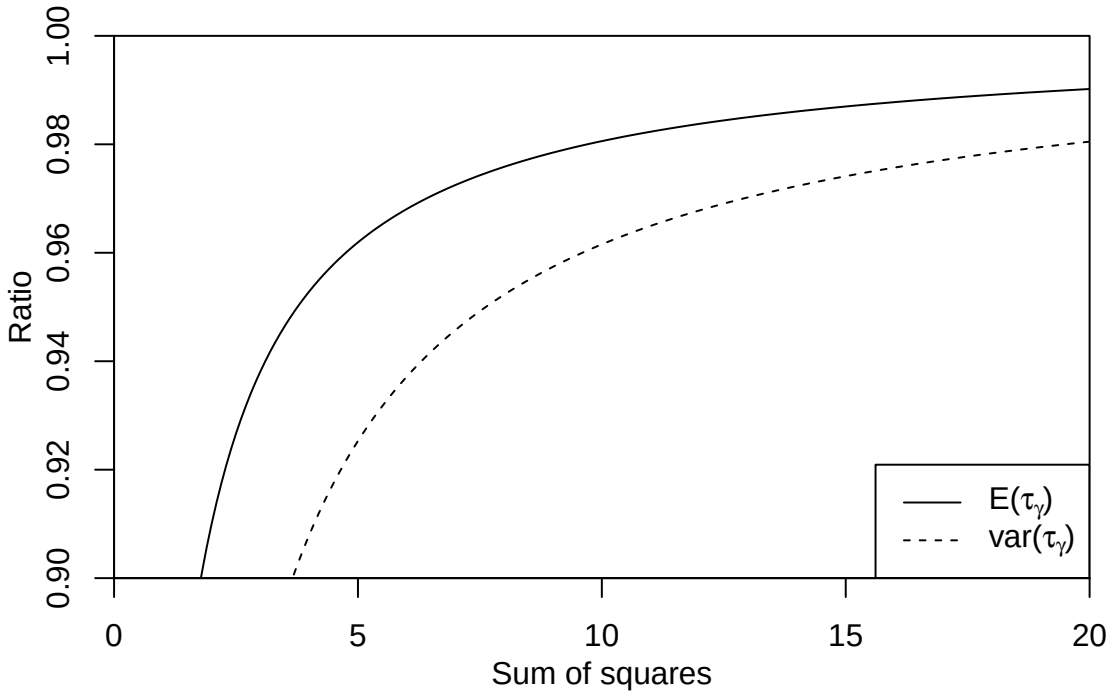

Figure 4.4.: The ratio of the conditional expectation and variance of τ_γ under the informative prior ($d_\gamma = 10^{-1}$) to those under the uninformative prior ($d_\gamma = 10^{-3}$) for various values of the sum of squared model response departures ($\sum_{m=1}^{M} \gamma_m^2$).

time series for τ_γ (not shown) also appears better mixed. Overall, τ_γ is still poorly determined, but its behaviour is no longer concerning. The estimates of μ, τ_H and τ_α are unaltered by the change in prior. When $d_\gamma = 10^{-1}$, the mode of the posterior density of τ_F increases and is more similar to that of τ_H. This is to be expected since a larger amount of variability is attributed to the model response departures by placing a lower bound on τ_γ. The additional variability in the model responses is reflected in the posterior density of β which broader for $d_\gamma = 10^{-1}$. The modified prior introduces only minimal prior information, solves a number practical difficulties, and explicitly reflects the belief that the models will never agree completely. Therefore it was adopted as the default prior for all grid points.

Based on the results for the modified prior, the estimate of β from the hierarchical framework is -0.05 $(-0.39, 0.29)$. The mean is unaltered from the two-way ANOVA estimate in Chapter 3, but the credible interval is slightly wider as a result of the additional uncertainty due to the model differences. The 95% credible interval for β under the original non-informative prior is $(-0.36, 0.25)$, more similar to the ANOVA interval. Using the ANOVA framework, the models were found to be in reasonable agreement about the climate response over London. Model agreement was defined as when the variability explained by the model response differences is small compared to that explained by the internal variability. The posterior credible intervals for β demonstrate exactly that. The increase in the width of the credible interval for β by allowing for model differences is small compared to width of the interval due only

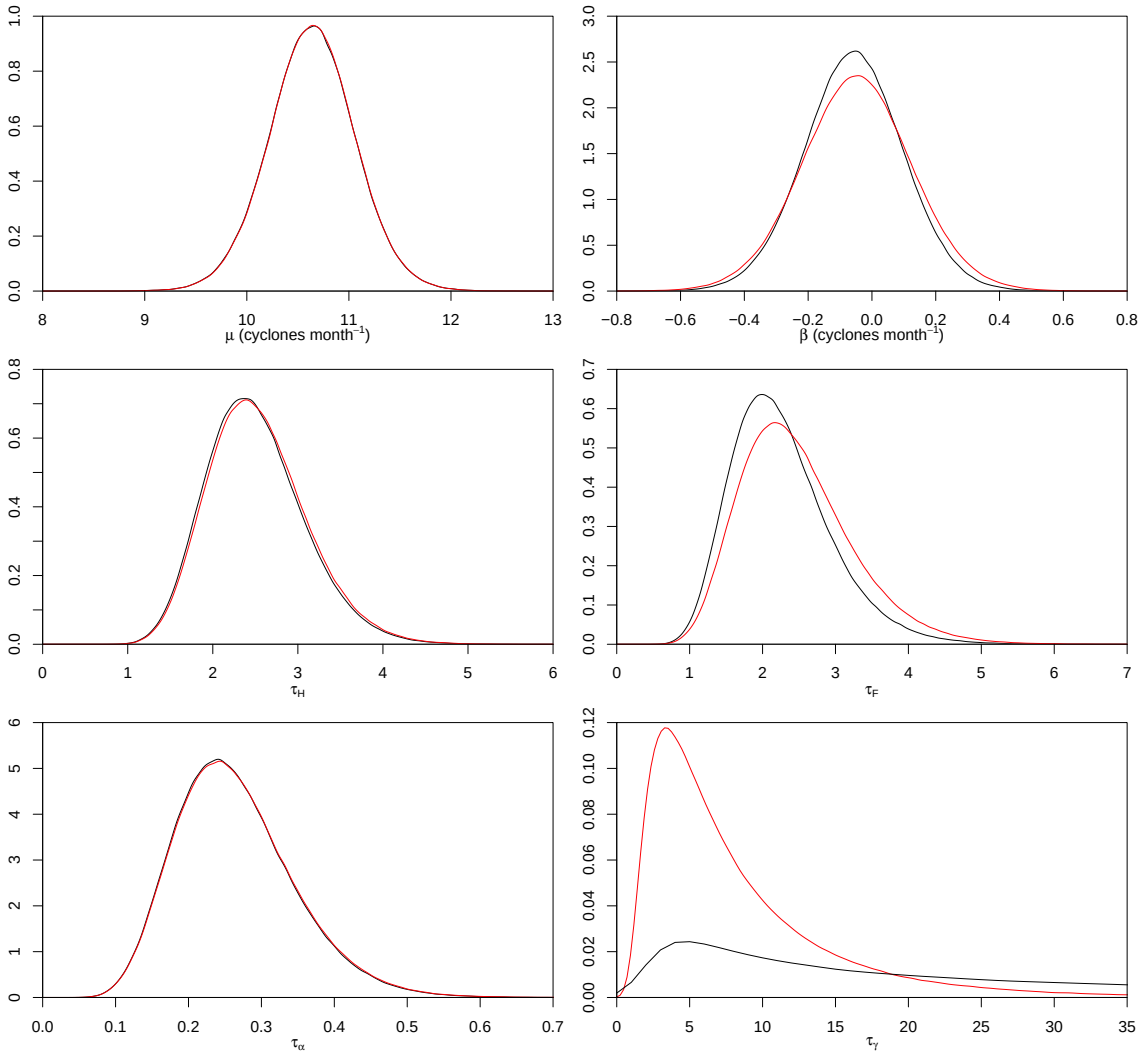

Figure 4.5.: Posterior densities of the parameters for the grid box containing London. Densities simulated with the uninformative prior $d_\gamma = 10^{-3}$ are shown in black. Densities simulated with the mildly informative prior $d_\gamma = 10^{-1}$ are shown in red.

to internal variability.

4.7.2. The North Atlantic storm track

After testing at a random selection of grid points, the hierarchical framework was fitted separately at each grid point in the cyclone track density data analysed in Chapter 3. The expected climate response of the models β is almost indistinguishable from the ANOVA estimates (Figure 4.6a). A comparison with Figure 3.4 suggests that the estimate from the hierarchical framework is most similar to the two-way ANOVA estimate, rather than the estimate from the ANOVA framework with interactions. In the hierarchical framework, the model responses are treated as an exchangeable sequence of random quantities. They are modelled as a random sample from a common distribution with expectation β. The expectation of the conditional distribution of β in Appendix B.1 is a weighted average, similar to the estimate from

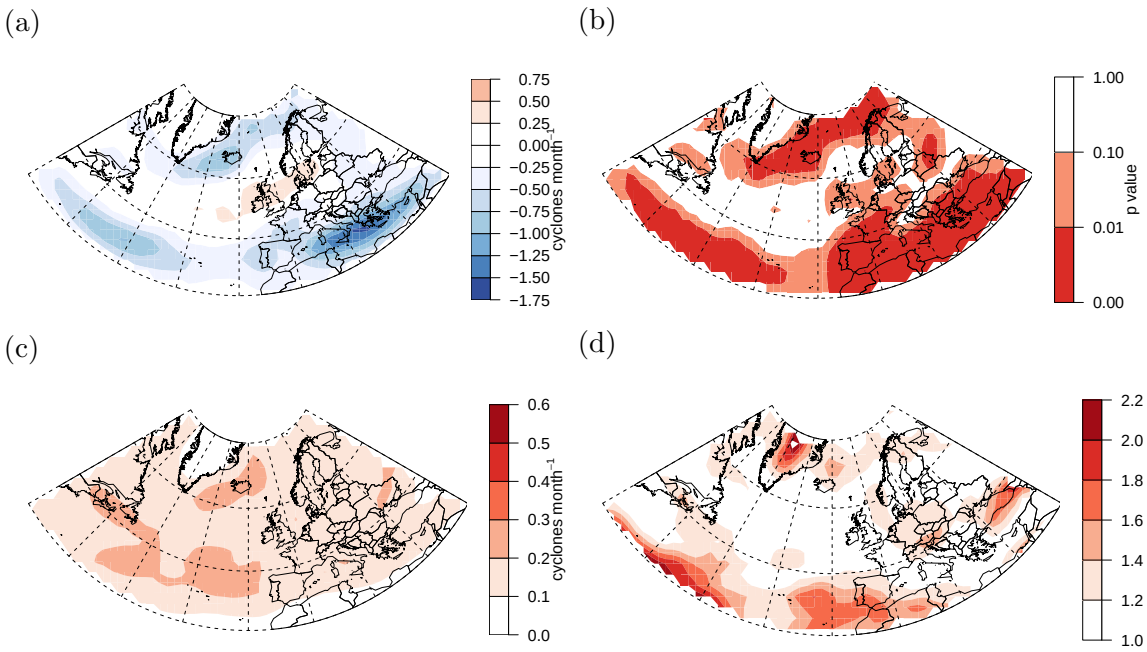

Figure 4.6.: (a) The posterior mean, (b) approximated p-value of the t test for non-zero climate response, and (c) the posterior standard error of the expected climate response β of the ensemble, estimated using the hierarchical framework; (d) the ratio of the posterior standard error of β to the standard error of β from the two-way ANOVA with interactions.

the two-way ANOVA, so the resemblance is not surprising. The pseudo p-value of the expected climate response β (Figure 4.6b) also strongly resembles the result of the t test under the two-way ANOVA framework.

The results in Chapter 3 indicated that the models were in good agreement about the climate response over most of the North Atlantic. Therefore, the contribution to the uncertainty about β from model differences should be small and so the posterior standard error and hence the p-value should be similar to the ANOVA estimates. The standard error from the two-way ANOVA framework will be inflated by the contribution from model differences where the models to not agree. Therefore, it is more instructive to compare the posterior standard error with the framework with interactions (Figure 4.6d). As expected, where there was good consensus on the model response between 45-60N in Chapter 3, the posterior standard error is very similar to the estimate from the ANOVA framework. However, on the southern flank of the storm track, the standard error is up to double that from the ANOVA framework.

The inter-model spread in the climate response σ_γ^2 is largest on the southern flank of the storm track where significant differences between the model responses were detected in Chapter 3. In this region, the square root of the inter-model spread may exceed 1.0 cyclone per month (Figure 4.7a). Where the models were found to agree on the climate response between 45-60N, the square root of the mean inter-

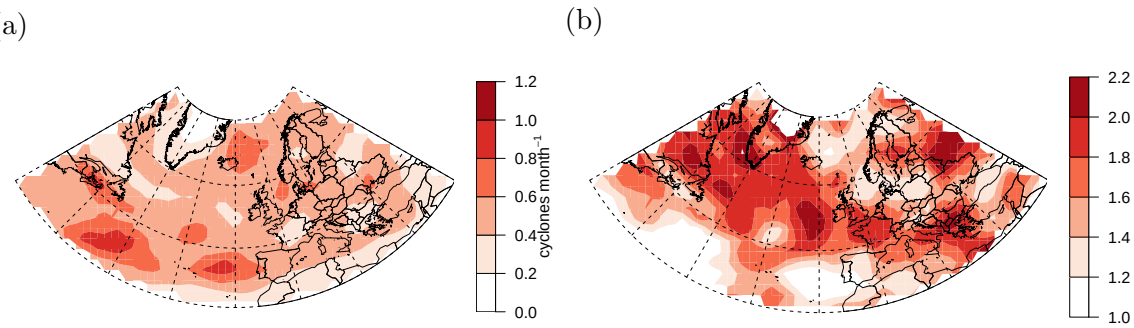

Figure 4.7.: (a) The square root of the posterior mean of the inter-model spread in the climate response σ_γ^2; (b) the ratio of (a) to the equivalent estimate using the uninformative parameterisation $d_\gamma = 10^{-3}$.

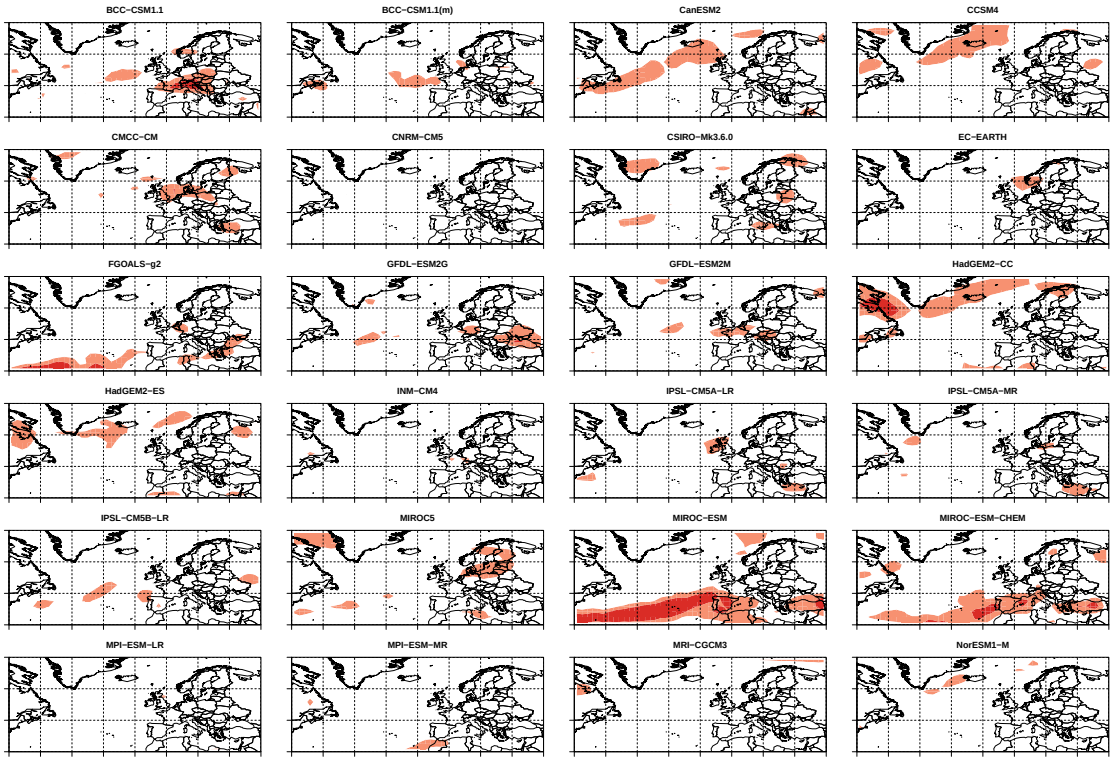

Figure 4.8.: p-values from the cross validation of the model mean climate responses, shading is the same as Figure 4.6b.

model spread usually falls between 0.4-0.6 cyclones per month. This is consistent with the mildly informative prior $d_\gamma = 10^{-1}$ constraining the inter-model spread to a minimum value. On the southern flank of the storm track the estimated inter-model spread is similar using either prior for τ_γ (Figure 4.7b). Between 45-60N where there is good consensus on the response according to the ANOVA formulation, the square root of the inter-model spread is up to two times greater than when estimated using the uninformative prior.

Cross-validation

The results of the cross-validation procedure on the climate response of each model are shown in Figure 4.8. The responses of three models, FGOALS-g2, MIROC-ESM and MIROC-ESM-CHEM, are poorly predicted on the southern flank of the storm track. Zappa et al. (2013a) identified all three of these models as belonging to a group of models whose storm tracks were displaced to the south. The responses of several models are also significant at the 10% level near Greenland and Iceland. This may also be attributable to the effect of the southward-displaced group influencing the expected response in this region. Although the model responses were found to agree in this region in Chapter 3, the internal variability is also large. It is possible that the model differences simply cannot be distinguished from the internal variability. This is explored further in the next section where the ensemble is thinned to obtain an exchangeable set of models.

Cross validation of the historical and future climates of the models (not shown) reveals no major cause for concern. However, FGOALS-g2 is poorly predicted in both scenarios in the most active part of the storm track, and IPSL-CM5A-LR which is poorly predicted at high latitudes. Overall, the hierarchical framework appears to provide a good description of the variation in extra-tropical cyclone frequency present in the CMIP5 ensemble.

Thinning the ensemble

In Section 4.3, the possibility was discussed that not all of the models should be included in the analysis in order to satisfy the assumption of exchangeability. In particular, models from the same centre are likely to be more similar than those from different centres. Therefore, only one model from each centre should be included. Modelling centres may also share components with other groups. Therefore, where possible only one model using any given major component, or at least any combination of components, should be included. These are judgements related to the structure of the model, however physical concerns may also be relevant.

Zappa et al. (2013a) identified three groups of models in terms of their bias relative to the storm track computed from the ERA-Interim reanalysis. Good performance in the historical scenario is no guarantee of good performance for the climate response. However, if a model simulates a physically implausible historical climate, then we may not be willing to judge that it is informative for the actual climate or exchangeable with the other models. The three groups identified by Zappa et al. (2013a) were (a) models compatible with ERA-Interim, (b) models whose storm tracks are too zonal and (c) models whose storm tracks are displaced to the south.

Table 4.1.: Number of realisations available from each model for the historical and future scenarios. Models highlighted in red are included in the exchangeable ensemble.

Modelling centre	Model	Runs		Storm Track Classification
		Historical	RCP4.5	
		R_{Hm}	R_{Fm}	
BCC	BCC-CSM1.1	3	1	Southward-displaced
BCC	BCC-CSM1.1(m)	1	1	Too zonal
CCCMA	CanESM2	5	1	Too zonal
NCAR	CCSM4	1	1	Too zonal
CMCC	CMCC-CM	1	1	Southward-displaced
CNRM-CERFACS	CNRM-CM5	5	1	Southward-displaced
CSIRO-QCCCE	CSIRO-Mk3.6.0	4	5	Southward-displaced
ICHEC	EC-EARTH	3	3	Consistent
LASG-CESS	FGOALS-g2	1	1	Southward-displaced
NOAA GFDL	GFDL-ESM2G	1	1	Too zonal
NOAA GFDL	GFDL-ESM2M	1	1	
MOHC	HadGEM2-CC	2	1	Consistent
MOHC	HadGEM2-ES	1	1	Consistent
INM	INM-CM4	1	1	Too zonal
IPSL	IPSL-CM5A-LR	4	4	
IPSL	IPSL-CM5A-MR	1	1	Too zonal
IPSL	IPSL-CM5B-LR	1	1	Southward-displaced
MIROC	MIROC5	5	3	Southward-displaced
MIROC	MIROC-ESM	3	1	Southward-displaced
MIROC	MIROC-ESM-CHEM	1	1	Southward-displaced
MPI-M	MPI-ESM-LR	3	3	Too zonal
MPI-M	MPI-ESM-MR	3	3	Too zonal
MRI	MRI-CGCM3	5	1	Consistent
NCC	NorESM1-M	3	1	Too zonal
Total		59 (25)	39 (15)	

Which group each model belongs to is identified in Table 4.1. Unfortunately, only four models in the full ensemble were judged to be compatible with ERA-Interim, and two of those are from the same centre. The North Atlantic storm track is often characterised as having two branches, one running North East past Iceland towards Norway, and one running more zonally towards Central and Northern Europe (Blender et al., 1997). Therefore we might be willing to judge that the models in the "too zonal" group are in fact exchangeable with the models in the "compatible" group, it is simply that their preferred state tends more to towards the zonal branch. However, it seems unlikely that the models in the "southward-displaced" group can be regarded as exchangeable with the models in the other two groups. Even if they are linearly displaced so that they simulate the "correct" response but in the "wrong" location, they will still bias the estimates unless the shift is incorporated into the estimation procedure. Therefore the "southward-displaced" group are judged *not* to be exchangeable with the other models and are excluded from the

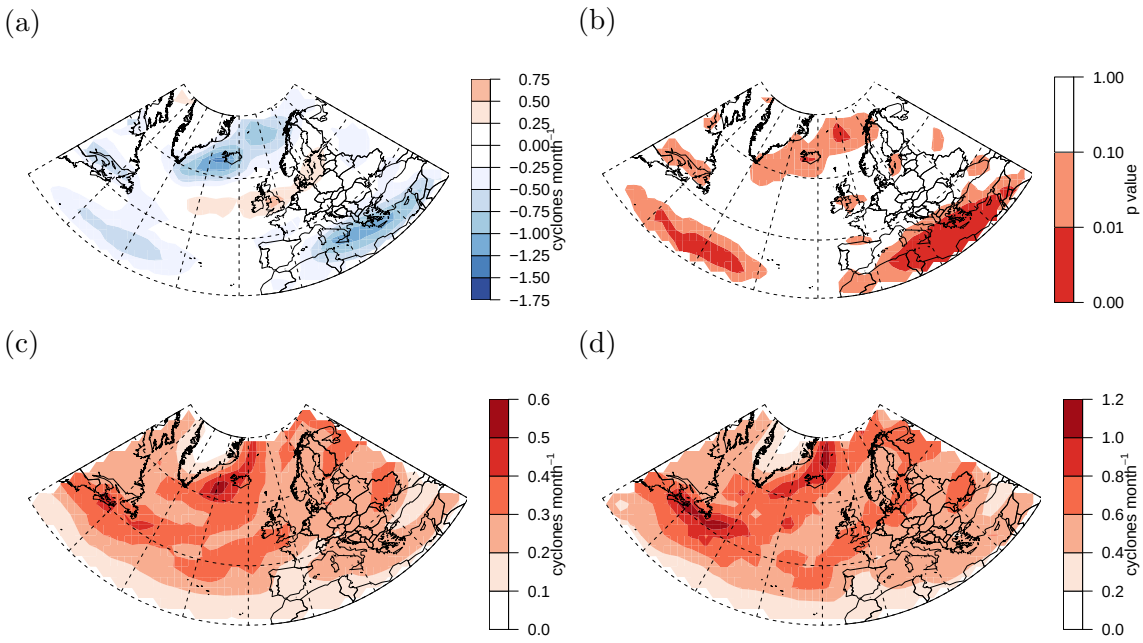

Figure 4.9.: (a) The posterior mean, (b) the approximated p-value of the t test for non-zero climate response, and (c) the standard error of the expected climate response (β); (d) the square root of posterior mean of the inter-model spread in the climate response ($\sqrt{\sigma_\gamma^2}$).

thinned ensemble.

After the "southward-displaced" models have been removed, 15 of 24 models remain. Among those, two were submitted from each of the NOAA-GFDL, MOHC, IPSL and MPI modelling centres. Structural details of all 24 climate models are summarised in Table 4.2. The two models from NOAA-GFDL differ primarily in their ocean component. GFDL-ESM2G uses the GOLD ocean model, while GFDL-ESM2M uses the MOM4.1 ocean model. However, the MOM4 ocean model is also used in the models from the BCC, so GFDL-ESM2M is excluded. The MOHC model in its HadGEM2-CC configuration has a relatively low resolution ocean component compared to most of the other models, so it is excluded in favour of the HadGEM2-ES configuration. Similarly, the resolution of the atmospheric component of the IPSL-CM5A-LR model is low compared to the rest of the ensemble, so it is excluded in favour of the IPSL-CM5A-MR configuration. Finally, the MPI-ESM-MR configuration features a very high resolution ocean component compared to the rest of the ensemble. Therefore the MPI-ESM-LR configuration is retained instead.

After thinning for both structural and physical concerns, an ensemble of 11 models remains. In total, there are 25 runs of the historical scenario, and 15 runs of the RCP4.5 scenario. With so few runs, it was anticipated that it might be necessary to reinstate the assumption of constant internal variability ($\sigma_F^2 = \sigma_H^2$), in order for the Markov chain to converge properly to the stationary distribution of the parameters. In practice however, the only problem observed was that the chain of samples for τ_α

Modelling centre	Model	Atmosphere	Atmosphere res.	Ocean	Ocean res.	Sea ice	Land surface
BCC	BCC-CSM1.1*	BCC_AGCM2.1	2.8 x 2.8 L26	MOM4-L40	1.0 x 1.0 L40	GFDL SIS	BCC-AVIM1.0
BCC	BCC-CSM1.1(m)	BCC_AGCM2.1	1.1 x 1.1 L26	MOM4-L40	1.0 x 1.0 L40	GFDL SIS	BCC-AVIM1.0
CCCMA	CanESM2	CanAM4	2.8 x 2.8 L35	CanOM4	1.4 x 1.4 L40	Included	CLASS 2.7; CTEM
NCAR	CCSM4	CAM4	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
CMCC	CMCC-CM*	ECHAM5	0.8 x 0.8 L31	OPA8.2	2.0 x 2.0 L31	LIM2	N / A
CNRM-CERFACS	CNRM-CM5*	ARPEGE-Climat	1.4 x 1.4 L31	NEMO	0.7 x 0.7 L42	Gelato5	SURFEX
CSIRO-QCCCE	CSIRO-Mk3.6.0*	Included	1.9 x 1.9 L18	MOM2.2	1.9 x 1.9 L31	Included	Included
EC-EARTH	EC-EARTH	IFS c3 1r1	1.1 x 1.1 L62	NEMO_ecmwf	1.0 x 1.0 L31	LIM2	HTESSEL
LASG-CESS	FGOALS-g2*	GAMIL2	2.8 x 2.8 L26	LICOM2	1.0 x 1.0 L30	CICE4-LASG	CLM3
NOAA GFDL	GFDL-ESM2G	AM2.1	2.5 x 2.5 L24	GOLD	1.0 x 1.0 L63	SIS	Included
NOAA GFDL	GFDL-ESM2M	AM2.1	2.5 x 2.5 L60	MOM4.1	1.0 x 1.0 L50	SIS	Included
MOHC	HadGEM2-CC	HadGAM2	1.9 x 1.9 L60	Included	1.9 x 1.9	Included	Included
MOHC	HadGEM2-ES	HadGAM2	1.9 x 1.9 L38	Included	1.0 x 1.0 L40	Included	Included
INM	INM-CM4	Included	2.0 x 2.0 L21	Included	1.0 x 1.0 L40	Included	Included
IPSL	IPSL-CM5A-LR	LMDZ5	3.8 x 3.8 L39	NEMO	2.0 x 2.0 L31	Included	Included
IPSL	IPSL-CM5A-MR	LMDZ5	2.5 x 2.5 L39	NEMO	2.0 x 2.0 L31	Included	Included
IPSL	IPSL-CM5B-LR*	LMDZ5	3.8 x 3.8 L39	NEMO	2.0 x 2.0 L31	Included	Included
MIROC	MIROC5*	MIROC-AGCM6	1.4 x 1.4 L40	COCO4.5	1.4 x 1.4 L50	Included	MATSIRO
MIROC	MIROC-ESM*	MIROC-AGCM	2.8 x 2.8 L80	COCO3.4	1.4 x 1.4 L44	Included	MATSIRO
MIROC	MIROC-ESM-CHEM*	MIROC-AGCM	2.8 x 2.8 L80	COCO3.4	1.4 x 1.4 L44	Included	MATSIRO
MPI-M	MPI-ESM-LR	ECHAM6	1.9 x 1.9 L47	MPIOM	1.5 x 1.5 L40	Included	JSBACH
MPI-M	MPI-ESM-MR	ECHAM6	1.9 x 1.9 L95	MPIOM	0.4 x 0.4 L40	Included	JSBACH
MRI	MRI-CGCM3	MRI-AGCM3.3	1.1 x 1.1 L48	MRI.COM3	1.0 x 1.0 L50	MRI.COM3	HAL
NCC	NorESM1-M	CAM4-Oslo	2.5 x 2.5 L26	NorESM-Ocean	1.1 x 1.1 L53	CICE4	CLM4

Table 4.2.: Structural details of the 24 CMIP5 models included in the analysis of cyclone track density. Models highlighted in red are included in the exchangeable ensemble. Models marked with an asterisk were excluded due to southward-displaced storm tracks. Atmosphere and ocean resolution are in degrees and Lxx indicates the number of vertical levels. Details included in this table were gathered from the metadata included in the model outputs and supplemented using information from Table 9.A.1 of Flato et al. (2013).

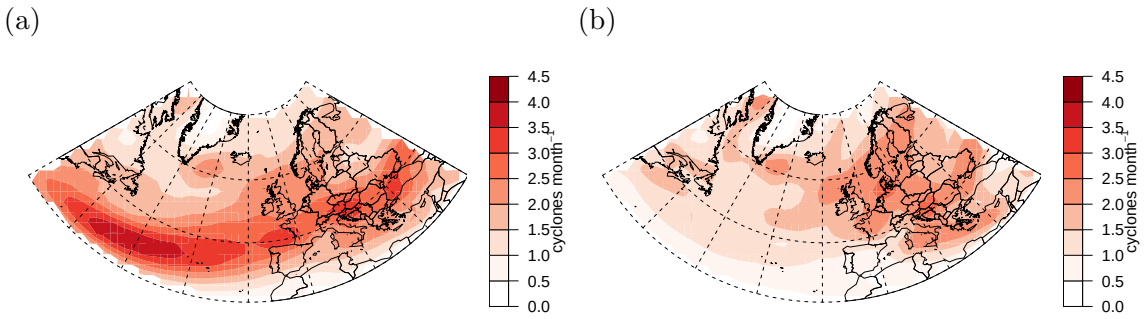

Figure 4.10.: The square root of the posterior mean of the inter-model spread in the historical climate ($\sqrt{\sigma_\alpha^2}$) estimated from (a) the full ensemble; and (b) the thinned ensemble.

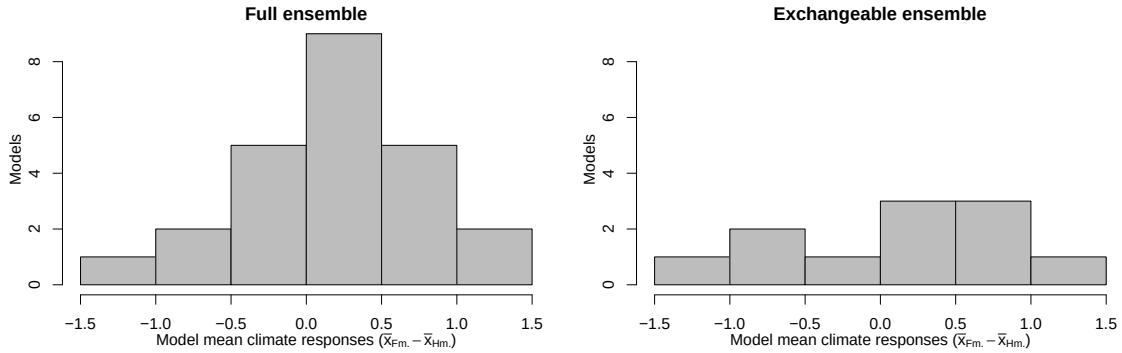

Figure 4.11.: Histograms of model mean climate responses for a grid box located off of Newfoundland (46.9W,51.6N).

displayed problems similar to those seen for τ_γ in the full ensemble. These issues were resolved using the same mildly informative prior for τ_α as for τ_γ (i.e., $d_\alpha = 10^{-1}$). The sampling issues for τ_α suggest that once the "southward-displaced" group are removed, the models are in good agreement about the historical climate in some regions. Figure 4.10 confirms that the inter-model spread in the historical scenario decreases dramatically in the sub-tropical North Atlantic in the thinned ensemble.

The expected response β of the thinned ensemble (Figure 4.9a) strongly resembles that of the full ensemble (Figure 4.6a). The response on the southern flank of the storm track and in the Mediterranean basin is weaker, but still present, despite the removal of the "southward-displaced" models. On the other hand, the response near Greenland and Iceland is strengthened. This is consistent with the speculation in the previous section that the "southward-displaced" models might be responsible for the difficulty in predicting the response of several models in the cross validation exercise. Cross-validation in the thinned ensemble (not shown) reveals no problems for the climate response or either individual scenario. This indicates that the hierarchical framework provides a good description of the thinned ensemble.

The standard error of the expected climate response β (Figure 4.9c) is increased considerably compared to the full ensemble (Figure 4.6c). This is to be expected,

since thinning the ensemble represents a significant reduction in information. While the increase in the standard error of β is expected, the increase and change in spatial distribution of the square root of the inter-model spread in the climate response $\sqrt{\sigma_\gamma^2}$ (Figure 4.9d) is perhaps more surprising. If the models were independent, then the inter-model spread should not be effected, although its estimate should be less precise. The histograms in Figure 4.11 illustrate how thinning the ensemble reveals the true variation between the models. In the full ensemble, the similarity between models from the same centres, and models that share components, cause the model mean climate responses to cluster between 0.0-0.5 cyclones per month. Thinning the ensemble removes the effect of the enhanced correlations between similar models. This leaves a much flatter histogram from which a larger inter-model spread would be estimated.

The increase in the standard error of the expected climate response β is due to a combination of the reduction in information and the increased inter-model spread in the response. As a result of the increased uncertainty, the significance of the expected climate response is reduced everywhere (Figure 4.9b). However, there is still a significant signal on the Northern flank of the storm track near Greenland and Iceland, in the Mediterranean basin and on the southern flank of the storm track south-west of the Azores, despite the removal of the "southward-displaced" group of models.

4.8. Discussion

In this chapter, the multi-model ensemble has been reinterpreted from a Bayesian perspective using the concept of exchangeability. However, climate scientists are already accustomed to making judgements about the exchangeability of climate models. They instinctively exclude older models as well as models that do not simulate key features in a physically plausible manner. The statistical concept of exchangeability provides a formal definition on which these judgements can be based. Questioning the exchangeability of the models in the CMIP5 ensemble leads to some interesting conclusions. In order for the conditions of exchangeability to be satisfied, each modelling centre should only submit output from one model. In addition, models from different centres should ideally not share major components. The extent to which the structural similarities of climate models are reflected in their outputs has been graphically illustrated by Masson and Knutti (2011) and Knutti et al. (2013). By thinning the ensemble, it appears that perfectly good outputs are being discarded. However, the analysis of the North Atlantic storm track in Section 4.7 showed that including many similar models can bias the estimates and lead to overconfidence. The same thinning should therefore be applied before fitting

the ANOVA frameworks derived in Chapter 3.

Many statistical frameworks for interpreting ensembles of climate models are based on only one run from each model (e.g., Buser et al., 2009; Smith et al., 2009) or the means of all the runs from each model (e.g., Bracegirdle and Stephenson, 2012; Greene et al., 2006). In either case, the inter-model spread will reflect not only the differences between the preferred climates of the models, but also a degree of uncertainty due to internal variability. Tebaldi et al. (2005) suggested incorporating the additional runs from each model by including model specific random effects. This would allow each model to simulate different internal variability. However, several runs of each scenario by each model would be required to estimate the individual uncertainties. The hierarchical framework described here uses the simplifying assumption that all models simulate the same internal variability. This allows uncertainty due to model differences and internal variability to be quantified separately.

By explicitly including model differences in the climate response, the hierarchical framework appears closest to the ANOVA framework with interactions. However, the assumption that the model responses are exchangeable implies that the models share a common mean response. The same is true for the historical climates of the models. So the hierarchical framework is in some respects closer to the two-way or even the one-way ANOVA framework. This can be seen from the expectations of the conditional distributions of μ and β (Equations B.4 & B.5). Like the two-way and one-way ANOVA estimates, these are weighted averages of the mean outputs of the individual models, and the models with the most runs in each scenario receive the most weight. This reinforces the conclusion in Chapter 3 that modelling centres need to submit several runs from *each* scenario in order for their model to receive a high weight.

In Equation 4.1, both the internal variability and the model departures are assumed to be normally distributed. Similar assumptions have been made in other studies, particularly for the model departures (e.g., Buser et al., 2009; Bracegirdle and Stephenson, 2012). In the absence of strong beliefs about the distribution of the departures, the symmetry of the normal distribution makes it a natural choice. However, the Bayesian implementation of the hierarchical framework is easily modified if we have reason to believe that the normal distribution is not appropriate. Tebaldi et al. (2005) chose a t distribution for the model departures. This places more weight in the tails of the distribution, making the estimates more robust to outlying models (Nychka and Tebaldi, 2003). The cost of modifying the normal assumption is that some of the full conditional distributions of the parameters may not take the form of any known probability distribution. In that case, the Metropolis-Hastings algorithm can be used to sample from the conditional posterior distribution of those parameters (Gelman et al., 2014).

The nested family of ANOVA frameworks derived in Chapter 3 correspond to different degrees of model consensus. A similar family of hierarchical frameworks could be derived for the same hypotheses. However, in Section 4.5.3 it was argued that in practice, the models should not be expected to agree completely. If required, the choice between a nested family of hierarchical frameworks, could be based on one of a number of information criteria. The best known is Akaike's Information Criteria (AIC) (Akaike, 1974), and a widely used Bayesian alternative is the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002). Superficially, the information criteria are likelihood ratio tests, similar to the F tests in Chapter 3, but penalised for model complexity. However, the justification for each is very different. For either criteria, the framework with the lowest value of the criteria should be preferred. Depending on the purpose of the analysis, there are reasons to favour either criteria (Spiegelhalter, 2006). One obstacle to performing framework selection in this way is the need to fit each framework in order to compute the value of information criteria. The ANOVA frameworks are simple to fit. However, iterative sampling from the hierarchical frameworks for a large number of grid points requires many hours of computer time which may be prohibitive.

Fitting the hierarchical framework to the full ensemble appeared to confirm the conclusions from ANOVA frameworks in Chapter 3. The contribution of model differences to the uncertainty about the expected change in cyclone frequency appeared to be small over most of the North Atlantic. However, thinning the ensemble revealed a very different picture. Qualitatively the pattern of the expected climate response in the storm track remains similar to that estimated in the full ensemble. However, the expected response near Greenland and Iceland is stronger in the thinned ensemble, while response in the sub-tropics and the Mediterranean basin is weaker. Near Greenland and Iceland, the standard error in the expected response of the models may be doubled in the thinned ensemble due to a combination of the smaller ensemble and the revised estimate of the model uncertainty.

4.9. Conclusion

In this chapter, it was shown that random effects can be used to quantify the uncertainty due to model differences about the climate response in an ensemble of climate models. In doing so, the ensemble was reinterpreted from a Bayesian perspective using the concept of exchangeability. The judgements required about climate model structure and the physical plausibility of the outputs are already familiar to climate scientists. When applied systematically to the CMIP5 ensemble, these judgements revealed a very different picture of model uncertainty in the North Atlantic storm track. This is not the first time that models have been treated as a population or

an exchangeable sequence. However, previous studies have usually included only a single run from each model or analysed the means of the runs from each model. The hierarchical framework presented here allows all the runs from each model to be entered into the estimation procedure. This has the advantage that the components of uncertainty due to internal variability and model differences can be quantified separately.

If it were assumed that the actual climate is exchangeable with the expected climates of the models, then the framework proposed here could be used to project the future climate response of the Earth system. However, in Chapter 2 it was argued that climate models were fundamentally different from the Earth system, due to inadequacies shared by all models. To naively include the observed climate as though it were the output of a model would also neglect uncertainty due to measurement error in the observations. Additional assumptions are required to account for the effects of model inadequacy and observation error. These will be addressed in Chapter 6. First, in Chapter 5, the hierarchical framework is extended to include the estimation of correlations between the historical climates and climate responses of the models, i.e., emergent constraints.

5. Incorporating emergent constraints

5.1. Introduction

The hierarchical framework described in Chapter 4 allows model uncertainty to be quantified in addition to the internal variability simulated by the models. In deriving that framework, it was assumed that the climate response of each model was independent of its historical climate. However, evidence was noted in Chapter 3 of a possible correlation between the responses of the CMIP5 models and their historical climates in the North Atlantic storm track. In the presence of such correlation, a systematic component is missing from the hierarchical framework and the model uncertainty about the climate response will be overestimated.

Correlations between the climate responses and historical states of the models are sometimes referred to as "emergent constraints" (Allen and Ingram, 2002). Simple linear regression methods are often used to characterise these relationships between the model climates (Räisänen et al., 2010; Cox et al., 2013; Karpechko et al., 2013). This methodology was formalised by Bracegirdle and Stephenson (2012) under the name "ensemble regression". Correlations between the responses and historical states of models have also been incorporated into more complex probabilistic frameworks for analysing multi-model ensembles (Tebaldi et al., 2005; Smith et al., 2009). However, those frameworks did not separate the effects of model uncertainty from those of internal variability.

In this chapter, the hierarchical framework described in Chapter 4 will be extended to include the effect of emergent constraints. It will be shown that it is important to allow for internal variability when estimating emergent relationships, in order to avoid biased estimates. The cross-validation methodology will also be extended to test the robustness and predictive value of the emergent constraint. Emergent constraints are primarily of interest for their potential to constrain projections of future climate. The emphasis in this chapter is still on building a probabilistic description of the ensemble itself. However, a brief discussion of projection is informative for framework checking and selection.

5.2. Extending the hierarchical framework

The ensemble regression methodology proposed by Bracegirdle and Stephenson (2012) is used to motivate the form of the extension to the hierarchical framework. In ensemble regression, the mean response of model m is modelled as linearly dependent on the mean historical climate of the same model

$$\overline{x}_{Fm.} - \overline{x}_{Hm.} = \beta + \lambda \left(\overline{x}_{Hm.} - \frac{1}{M} \sum_{m=1}^{M} \overline{x}_{Hm.} \right) + \epsilon_m \tag{5.1a}$$

$$\epsilon_m \overset{iid}{\sim} N\left(0 \ , \ \sigma^2\right) \tag{5.1b}$$

Where β represents the expected response of the ensemble as usual, and the slope parameter λ represents the emergent constraint. The variance σ^2 quantifies the inter-model spread in sample mean responses. Note that Bracegirdle and Stephenson (2012) treat the sample means $\overline{x}_{Hm.}$ as fixed quantities, measured without error. However, probabilistic descriptions of the individual runs x_{smr} were specified in the hierarchical framework developed in Chapter 4. The weak law of large numbers guarantees that the sample means will converge to the expected climates and climate responses of the models defined in the hierarchical framework.

$$\begin{aligned}
\overline{x}_{Hm.} &\to \mu + \alpha_m && \text{as} && R_{Hm} \to \infty \\
\frac{1}{M} \sum_{m=1}^{M} \overline{x}_{Hm.} &\to \mu && \text{as} && M, R_{Hm} \to \infty \\
\overline{x}_{Fm.} - \overline{x}_{Hm.} &\to \beta + \gamma_m && \text{as} && R_{Hm}, R_{Fm} \to \infty
\end{aligned}$$

The maximum likelihood estimate of the expected climate response in ensemble regression

$$\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \left(\overline{x}_{Fm.} - \overline{x}_{Hm.} \right)$$

will also converge to β as the ensemble size increases. In the limit as the number of models and runs becomes large, Equation 5.1a is asymptotically equivalent to

$$\beta + \gamma_m = \beta + \lambda \left(\mu + \alpha_m - \mu \right) + \epsilon_m$$

which after cancelling the μ and β terms leaves

$$\gamma_m = \lambda \alpha_m + \epsilon_m \tag{5.2}$$

i.e., the expected departure of model m from the expected climate response β, is proportional to its historical departure from the expected climate μ. This suggests

the following extension to the hierarchical framework

$$x_{Hmr} \overset{iid}{\sim} N\left(\mu + \alpha_m \ , \ \sigma_H^2\right) \tag{5.3a}$$

$$x_{Fmr} \overset{iid}{\sim} N\left(\mu + \alpha_m + \beta + \gamma_m \ , \ \sigma_F^2\right) \tag{5.3b}$$

$$\alpha_m \overset{iid}{\sim} N\left(0 \ , \ \sigma_\alpha^2\right) \tag{5.3c}$$

$$\gamma_m \mid \alpha_m \overset{iid}{\sim} N\left(\lambda\alpha_m \ , \ \sigma_{\gamma|\alpha}^2\right) \tag{5.3d}$$

which is equivalent to the basic hierarchical framework in Chapter 4 when $\lambda = 0$.

Since Equation 5.3 represents a relatively minor generalisation of the basic hierarchical framework, the assumptions and interpretations of the parameters are largely the same as in Chapter 4. The only changes relate to the model response departures γ_m, which now depend linearly on the historical departures α_m. The magnitude of the dependence is controlled by the emergent constraint λ. Note the change in notation from σ_γ^2 to $\sigma_{\gamma|\alpha}^2$ for the inter-model spread in the response. σ_γ^2 is the marginal variance of the expected model response and $\sigma_{\gamma|\alpha}^2$ is the conditional variance, given the expected historical climate of the model.

5.2.1. Fitting the extended framework

In order to fit the extended hierarchical framework, it is necessary to specify a prior probability distribution for λ. Once again, a diffuse normal prior is chosen

$$\lambda \sim N\left(a_\lambda, b_\lambda^{-1}\right) \tag{5.4}$$

where $a_\lambda = 0$ and $b_\lambda = 10^{-6}$. The priors on the other parameters are the same as for the basic hierarchical framework in Chapter 4. The full conditional distributions of the parameters are derived in Appendix C. The joint posterior distribution can be approximated by Gibbs sampling. The maximum likelihood estimate from ensemble regression is used as an intial value for the emergent constraint λ. A sample estimate is used to initialise the conditional precision of the model response differences $\tau_{\gamma|\alpha} = \sigma_{\gamma|\alpha}^{-2}$.

$$\tau_{\gamma|\alpha} = \frac{M-1}{\sum_{m=1}^{M}\left(\gamma_m - \lambda\alpha_m\right)^2} \tag{5.5}$$

The initial values for the other parameters are unchanged from the basic hierarchical framework in Chapter 4

5.3. Interpreting emergent relationships

The form of the extended hierarchical framework highlights an important assumption about the nature of emergent relationships. In Equation 5.3, the emergent constraint λ relates the differences between the models α_m and γ_m, but not the differences between the individual runs. This reflects the usual assumption that the trajectory of the climate over long time scales is insensitive (i.e., independent) to fluctuations due to internal variability on short time scales. For example, Boé et al. (2009) linked differences in the thickness distribution of sea ice simulated by the CMIP3 models in the historical period to differences in their percentage ice loss in the future. Suppose model j simulates more thin ice during the historical period in a particular run, compared to its expected (preferred) ice thickness distribution. Thin ice is melted more easily than thick ice. However, we would not expect that the same run would necessarily have less sea ice remaining in the future period compared to the expected future ice coverage in that model. The anomaly in the historical period can be thought of as the accumulation of year-to-year variations that are assumed to have no effect on the long term trajectory of the climate. Now, suppose that model k systematically simulates more thin ice than model j in the historical period, i.e., its ice thickness distribution is biased. In a warming climate, we would rightly expect that model k will tend to have less ice remaining in the future compared to model j.

5.3.1. Why internal variability matters

Most previous studies of emergent relationships have utilised only one run from each model (e.g., Boé et al., 2009; Hall and Qu, 2006; Tebaldi et al., 2005). Those studies did not attempt to distinguish between model differences and internal variability. If the internal variability is large compared to the model departures, then it may be impossible to detect any systematic relationship between the models. Bracegirdle and Stephenson (2012) reduce the impact of internal variability by averaging together all the runs from each model. However, this can actually result in the estimation of an emergent relationship where none exists. Consider a balanced ensemble where each model simulates the same number of runs of each scenario (i.e., $N_{Hm} = N_{Fm} = N$). The covariance between the sample mean responses and historical climates of the

models is

$$\begin{aligned}\operatorname{cov}\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\ ,\ \overline{x}_{Hm.} \mid \mu, \beta, \lambda\right) &= \operatorname{cov}\left(\beta + \gamma_m + \sum_{r=1}^{N} \frac{\varepsilon_{Fmr}}{N} - \sum_{r=1}^{N} \frac{\varepsilon_{Hmr}}{N}\ ,\ \mu + \alpha_m + \sum_{r=1}^{N} \frac{\varepsilon_{Hmr}}{N}\right) \\ &= \operatorname{cov}\left(\alpha_m\ ,\ \gamma_m\right) - \frac{\operatorname{cov}\left(\varepsilon_{Hmr}\ ,\ \varepsilon_{Hmr}\right)}{N} \\ &= \lambda\sigma_\alpha^2 - \frac{\sigma_H^2}{N} \end{aligned} \tag{5.6}$$

The maximum likelihood estimate of the emergent constraint in ensemble regression can be written in terms of the sample variance and covariance of the model mean climates

$$\hat{\lambda} = \frac{\operatorname{cov}\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\ ,\ \overline{x}_{Hm.}\right)}{\operatorname{var}\left(\overline{x}_{Hm.}\right)} \tag{5.7}$$

the expected value of which is

$$\mathrm{E}\left(\hat{\lambda}\right) = \frac{\lambda\sigma_\alpha^2 - \sigma_H^2/N}{\sigma_\alpha^2 + \sigma_H^2/N} = \lambda - \frac{\left(\lambda + 1\right)\sigma_H^2/N}{\sigma_\alpha^2 + \sigma_H^2/N} \tag{5.8}$$

Because of the shared term $\overline{x}_{Hm.}$, the ensemble regression estimate is negatively biased unless $\lambda = -1$ or $\sigma_H^2 = 0$, i.e., there is no internal variability. In particular, if there is no emergent constraint ($\lambda = 0$), then ensemble regression will tend to estimate an emergent relationship with negative sign. In fact, if $\lambda = 0$ and $N = 1$, then $\hat{\lambda} = -\sigma_H^2/\left(\sigma_\alpha^2 + \sigma_H^2\right)$, and $\hat{\lambda} \to -1$ in the limit as $\sigma_H^2/\sigma_\alpha^2 \to +\infty$. Therefore, additional care must be taken when applying the basic ensemble regression method of Bracegirdle and Stephenson (2012) when the internal variability is large and the number of runs from each model is small. The extended hierarchical framework in Equation 5.3 accounts for the uncertainty about the expected climates of the models. As a result, it is not affected by this bias unless there is only one run from each model $N = 1$, when the internal variability cannot be estimated.

5.4. Inference in the extended framework

Point estimates and credible intervals can be constructed as described in Chapter 4. In addition to the expected climate response of the ensemble β, the value of the emergent constraint λ will be of particular interest. In the ensemble regression framework of Bracegirdle and Stephenson (2012), a two-sided t test could be used to test the null hypothesis of no emergent relationship, $H_0 : \lambda = 0$, against the alternative hypothesis of any relationship, $H_\lambda : \lambda \neq 0$. The p-value can be approximated in the same way as described for the test for non-zero climate response in Chapter 4,

by computing

$$2 \times \min\left(\Pr\left(\lambda > 0\right)\ ,\ 1 - \Pr\left(\lambda > 0\right)\right) \tag{5.9}$$

i.e., the probability that the value of λ is more extreme than 0.

5.4.1. Prediction

Prediction of new runs is almost unchanged from Chapter 4. The posterior predictive distribution of the response departure $\tilde{\gamma}_j$ of a new model j in Equation 4.12 is altered by the inclusion of the emergent constraint. Its new predictive distribution can be sampled from

$$\tilde{\gamma}_j^{(n)} \mid \mathbf{x} \sim N\left(\lambda^{(n)}\tilde{\alpha}_j^{(n)}\ ,\ \sigma_\gamma^{2\,(n)}\right) \tag{5.10}$$

for each of the $n = 1, \ldots, N$ samples from the posterior distribution of the parameters. The posterior predictive distribution of the difference between a future and a historical run from new model j (Equation 4.14) is also altered, and can be sampled directly from

$$\widetilde{x_{Fjr} - x_{Fjr'}}^{(n)} \mid \mathbf{x} \sim N\left(\beta^{(n)}\ ,\ {\lambda^{(n)}}^2\sigma_\alpha^{2\,(n)} + \sigma_\gamma^{2\,(n)} + \sigma_H^{2\,(n)} + \sigma_F^{2\,(n)}\right) \tag{5.11}$$

5.4.2. Predicting the actual climate

The conditional variance of the expected model responses $\sigma^2_{\gamma|\alpha}$ is related to the marginal variance σ^2_γ by the law of total variance variance

$$\sigma^2_\gamma = \lambda^2\sigma^2_\alpha + \sigma^2_{\gamma|\alpha} \tag{5.12}$$

Clearly $\sigma^2_{\gamma|\alpha} < \sigma^2_\gamma$ for $\lambda \neq 0$, i.e., emergent relationships act to reduce the uncertainty about the expected response of a model, given its expected historical state. This effect has been exploited in order to reduce uncertainty about the response of the actual climate given knowledge of the historical climate from observations (Bracegirdle and Stephenson, 2012; Cox et al., 2013; Karpechko et al., 2013). By estimating the uncertainty about the actual climate response by the uncertainty about the expected response of a new model, it is implicitly assumed that the actual climate is exchangeable with expected climates of the models. In Chapter 2, it was argued that this assumption may be too strong since the models are unlikely to explore the full extent of our structural uncertainty and there are processes not represented in any climate model. A less restrictive alternative is proposed in Chapter 6.

5.4.3. Framework checking

The checks for convergence and autocorrelation described in Chapter 4 can be used to select appropriate burn-in and thinning strategies for the extended hierarchical model. The cross-validation procedures described previously can also be used to test the assumptions about the marginal distributions of the model climates and climate responses. However, additional checks are required for the emergent constraint and the conditional distribution of the model responses. The emergent relationship is modelled by a linear regression. Linear regressions can be influenced by data points that are outlying in both the response and explanatory variables, i.e., models that have very different climate responses *and* historical climates compared to the rest of the ensemble. Simply plotting the model mean climate responses ($\overline{x}_{Fm.} - \overline{x}_{Hm.}$) against the model mean historical climates ($\overline{x}_{Hm.}$) should reveal any potentially influential models. If any influential models are found, then the extended framework could be refitted without those models and the results compared. If the estimate of the emergent constraint changes dramatically, then it might be sensible to question whether or not to include the emergent constraint in the model, or whether those models are in fact exchangeable with the rest of the ensemble. As discussed in Chapters 3 and 4, the decision to remove a model from the ensemble *completely* should not be made simply because it differs from the other models. Such a decision should be based on expert judgement about its structure and the plausibility of its output.

Checking the emergent relationship by removing influential models suggests an extension to the cross validation approach. Rather than removing each model completely and then predicting its response, we could instead remove only the future runs, and then predict their mean (or the mean response) conditional on the historical runs. If a particular model either does not conform to the emergent relationship or is influential in inducing the relationship, then its response will be poorly predicted. The full conditional distributions of the parameters given the N_{Hj} historical runs from model j but excluding the N_{Fj} future runs are derived in Appendix C.2. The posterior distribution of α_j is estimated along with all the other historical departures. Therefore the posterior predictive distribution of the mean of the N_{Fj} future runs from model j can be approximated by taking one sample from

$$\widetilde{\overline{x}_{Fj.}}^{(n)} \mid \mathbf{x}_{m\neq j}, \mathbf{x}_{Hj} \sim N\left(\mu^{(n)} + \alpha_j^{(n)} + \beta^{(n)} + \lambda^{(n)}\alpha_j^{(n)} \ , \ {\sigma^2_{\gamma|\alpha}}^{(n)} + \frac{{\sigma_F^2}^{(n)}}{N_{Fj}}\right) \quad (5.13)$$

for each of the N samples of the posterior distribution of the parameters. Here $\mathbf{x}_{m\neq j}$ denotes the runs of all the models except model j and $\mathbf{x}_{Hj}$ denotes the historical runs of models j. Since the historical runs of model j are included in the estimation of the parameters, their mean $\overline{x}_{Hj.}$ is known. Therefore the posterior predictive

distribution of the mean response of model j is simply

$$\Pr\left(\overline{x}_{Fj.} - \overline{x}_{Hj.} \mid \mathbf{x}_{m\neq j}, \mathbf{x}_{Hj}\right) = \Pr\left(\overline{x}_{Fj.} \mid \mathbf{x}_{m\neq j}, \mathbf{x}_{Hj}\right) - \overline{x}_{Hj.} \tag{5.14}$$

and so conditional cross validation of the mean response is precisely equivalent to conditional cross validation of the future mean climate. Cross validation proceeds as in Chapter 4

1. For each $j \in 1, \ldots, M$ refit the hierarchical framework, leaving out the historical runs from model j;

2. For each of the N samples from the new posterior distribution of the parameters, draw one sample from the posterior predictive distribution of the sample mean future climate of model j (Equation 5.13);

3. Calculate the proportion of samples from the posterior predictive distribution that are greater than or equal to the value of the sample mean future climate of model j

$$p_j = \frac{1}{N}\sum_{n=1}^{N} \mathrm{I}\left(\widetilde{\overline{x}_{Fj.}}^{(n)} > \overline{x}_{Fj.}\right) \tag{5.15}$$

If the posterior predictive p-value is significant at the 1% level ($p_j \leq 0.005$ or $0.995 \leq p_j$) then model j should be examined carefully for inconsistent behaviour and possibly excluded from the ensemble. If several models have significant posterior predictive p-values, then the distributional assumptions about the model departures or the inclusion of the emergent constraint may need to be reassessed.

Bracegirdle and Stephenson (2013) suggest automating the check for influential models by calculating the area averaged Cook's distance of each model. Cook's distance is equivalent to cross validation for linear regression frameworks (Krzanowski, 1998). A large Cook's distance indicates that a model is influential and should be investigated further. Fitting the ensemble regression framework and calculating Cook's distance is computationally cheap compared to refitting the extended hierarchical framework for each model j. Since the ensemble regression method is essentially embedded within the extended hierarchical framework, evaluating Cook's distance may be used as an approximation to cross-validation when there are many grid points to be checked.

5.4.4. Framework selection

In Chapter 4 it was argued that framework selection was unnecessary in the basic hierarchical framework, since the climate model outputs are never expected to agree

completely. The situation is more complicated when emergent constraints are considered. If there is no emergent relationship, then the posterior distribution of λ should be concentrated near 0. In that case, the only cost of including the emergent constraint is a slight increase in the uncertainty of the other parameters.

In this chapter, the focus is on building a probabilistic description of the climate model outputs. However, emergent relationships are primarily of interest for their potential to constrain projections of the actual future climate. Knutti et al. (2010b) note that if there is no emergent relationship, then estimating one "will not constrain prediction but may introduce spurious biases". Simple graphical checks and cross-validation exercises should be sufficient to detect spurious relationships caused by outlying models (Section 5.4.3). The approximated p-value of λ and the information criteria described in Chapter 4 are useful indicators that a detected relationship did not occur by chance. Unfortunately, it is not possible to determine whether or not any detected relationship corresponds to a "real" physical process using only evidence from the models. The decision to include an emergent constraint for the purpose of projection must ultimately rely on expert judgement. If the physical processes underlying a relationship between the models are not well understood, then projections with and without the emergent constraint should be presented and contrasted.

Extra care must be taken when interpreting emergent relationships estimated independently over many grid boxes. In that case, strong relationships are expected to occur by chance in a small number of cases. Correlations between grid boxes mean that spatially contiguous regions with significant emergent constraints should not necessarily increase our confidence that the relationship is "real". However, if the relationship is not contiguous in space, then it is questionable whether it corresponds to any physical process.

5.5. Application to the North Atlantic storm track

In this section, the extended hierarchical framework and conditional cross validation methodology are applied to the North Atlantic storm track data from the previous chapters. According to Equation 4.19, the expected increase in runtime over the basic hierarchical framework is only 1.8%. This holds true for the Fortran implementation, which takes around 8.2 seconds to generate $N = 10^6$ samples from the joint posterior ($M = 24$), an increase of approximately 2.5%. The R implementation on the other hand takes around 182 seconds, an increase of approximately 12.5%. This kind of non-linear behaviour can occur in interpreted languages such as R, that

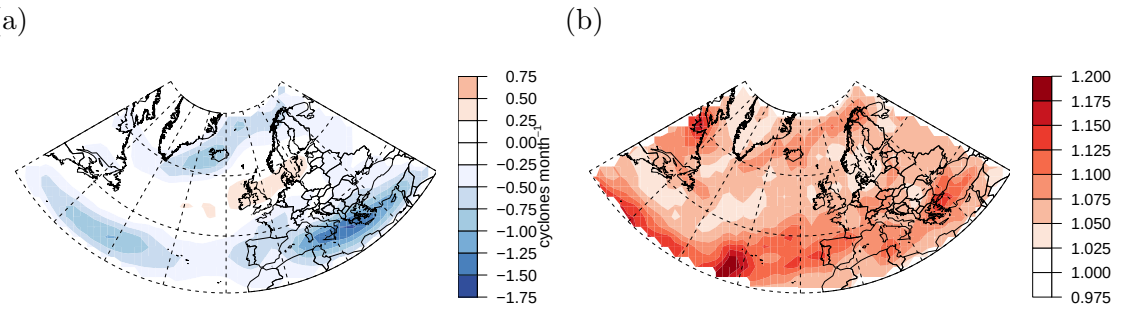

Figure 5.1.: (a) The posterior mean of the expected climate response of the ensemble (β), estimated using the extended hierarchical framework; (b) the ratio of the posterior standard error of β from the extended framework to the standard error the basic framework (no emergent constraint).

are unable to properly optimise either loops or memory management.

The mildly informative priors on the inter-model spread in the historical climate and climate response ($d_\alpha = d_\gamma = 10^{-1}$) were carried over from Chapter 4. All other prior probabilities remain unchanged and no further problems were noted. Checks for convergence and autocorrelation were carried out at a number of grid points across the study region. Introducing the emergent constraint increases the length of the autocorrelations. The samples from the expected historical climate μ may exhibit significant autocorrelation for up to 500 samples (not shown). Therefore the thinning strategy was revised so that only every 500th sample was retained. As before, little or no burn-in period was evident, the chains stabilised very quickly. However, the burn-in period was increased to 5,000 samples in order to allow for the increased autocorrelation.

The results for the grid point containing London are unremarkable. Without the emergent constraint, the 95% credible interval for the expected response β was -0.05 $(-0.39, +0.29)$. With the emergent constraint included, the interval is -0.03 $(-0.40, +0.33)$. The mean is essentially unchanged. There is a small increase in the width of the credible interval due to the estimation of the additional parameter. The 95% credible interval for the value of the emergent constraint λ is 0.06 $(-0.14, +0.26)$, so there is no significant evidence of an emergent relationship. If there was strong evidence of a relationship, then we should expect a reduction in the estimated inter-model spread in the climate response ($\sigma^2_{\gamma|\alpha} < \sigma^2_\gamma$). However, no relationship is evident and so the posterior mean and credible interval for $\sigma^2_{\gamma|\alpha}$ are identical to the estimates of σ^2_γ in Chapter 4. The estimates of the other parameters are also unchanged.

As in Chapter 4, the extended hierarchical framework was fitted grid-box by grid-box to the track density data from the full ensemble in the North Atlantic domain. The posterior mean of the expected climate response β is indistinguishable from that estimated by the basic hierarchical framework (Figure 5.1a). The standard error of

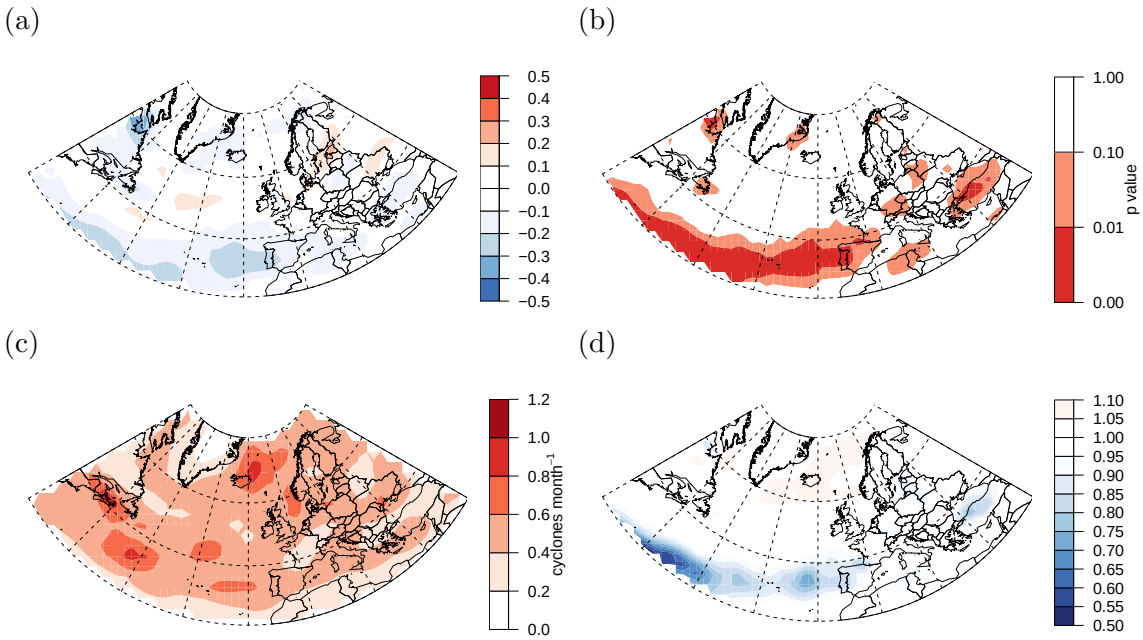

Figure 5.2.: (a) The posterior mean of the emergent constraint λ, (b) the approximated p-value of the t test for non-zero emergent constraint, (c) the square root of the posterior mean of the conditional inter-model spread in the response $\sqrt{\sigma^2_{\gamma|\alpha}}$, and (d) the ratio of (c) to the estimate of the marginal inter-model spread ($\sqrt{\sigma^2_{\gamma}}$)from the basic hierarchical framework with no emergent constraint.

β increases by a small amount across the study region due to the estimation of the additional parameter (Figure 5.1b).

The correlation between the model responses and historical states in the sub-tropics noted in Chapter 3 is captured by the emergent constraint λ (Figure 5.2a). In general, the more storms a model simulates in the historical scenario, the fewer storms it will simulate in the future scenario between 30N-45N. The approximated p-value of the t test for no emergent constraint suggests that the posterior probability of no relationship in that region is small (Figure 5.2b). The conditional inter-model spread in the response $\sigma_{\gamma|\alpha}$ (Figure 5.2c) is similar to the marginal estimate σ_{γ} (Figure 5.2c) over most of the study region. However, the conditional spread is decreased where there is strong evidence of an emergent relationship (Figure 5.2d), in agreement with Equation 5.12.

The maximum likelihood estimate of the emergent constraint λ from ensemble regression is generally stronger (more negative) than the estimate from the extended hierarchical framework (Figure 5.3b). This agrees with the theoretical arguments in Section 5.3. Ensemble regression suggests the presence of an emergent relationship in the main storm track near Iceland Figure 5.3a). The internal variability in this region is large (not shown) so the bias in the ensemble regression estimate is also expected to be large. A closer inspection suggests that four models are influencing the estimate. CCSM4, HadGEM2-CC, HadGEM2-ES and NorESM1-M tend to

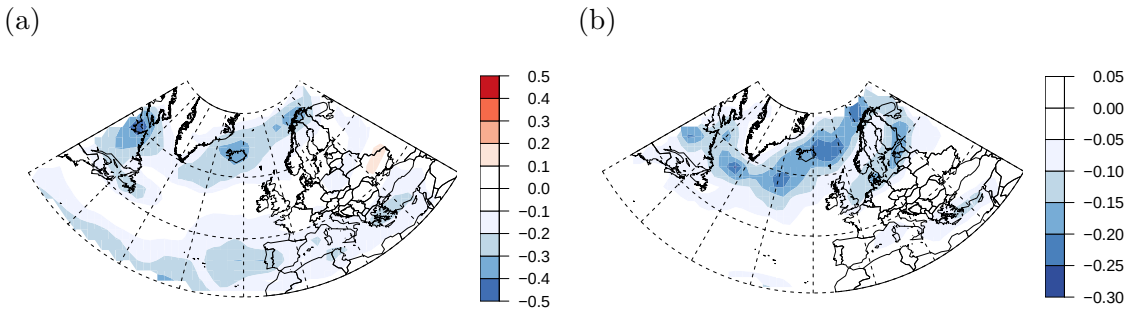

Figure 5.3.: (a) The maximum likelihood estimate of the emergent constraint λ from ensemble regression, and (b) the difference between (a) and the estimate of λ from the extended hierarchical framework (Figure 5.2a).

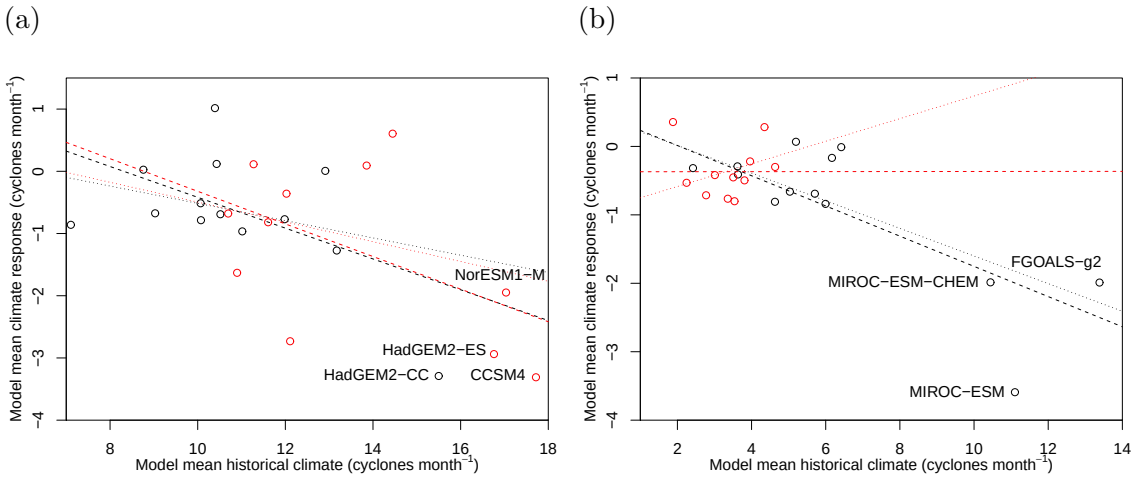

Figure 5.4.: The model mean climate responses ($\overline{x}_{Fm.} - \overline{x}_{Hm.}$) plotted against the model mean historical climates ($\overline{x}_{Hm.}$) for grid points (a) between Greenland and Iceland (36.8W,61.6N), and (b) near the Azores (29.3W,36.5N). Red points indicate models that are included in the exchangeable ensemble. Dashed lines represent the emergent relationships estimated by ensemble regression. Dotted lines are estimated using the extended hierarchical framework. Black lines are estimated from the full ensemble. Red lines are estimated from the exchangeable ensemble.

simulate very active storm tracks in the historical scenario and strong responses in the future scenario (Figure 5.4a). However, CCSM4 and HadGEM2-ES have only one run of the historical scenario, and HadGEM2-CC has only two runs. All of these models have only one run of the future scenario. With so few runs, the uncertainty about the expected climates of these models is large. The hierarchical framework accounts for this uncertainty and so is not influenced as strongly by these outlying runs.

When the emergent constraint is estimated from the exchangeable ensemble rather than the full ensemble, no systematic negative correlation is evident in the subtropics (Figure 5.5a). A weak positive correlation is visible slightly to the north of the negative correlation in the full ensemble. This suggests that one or more models may also be influencing the fit of the emergent constraint south of 30N in the full ensemble.

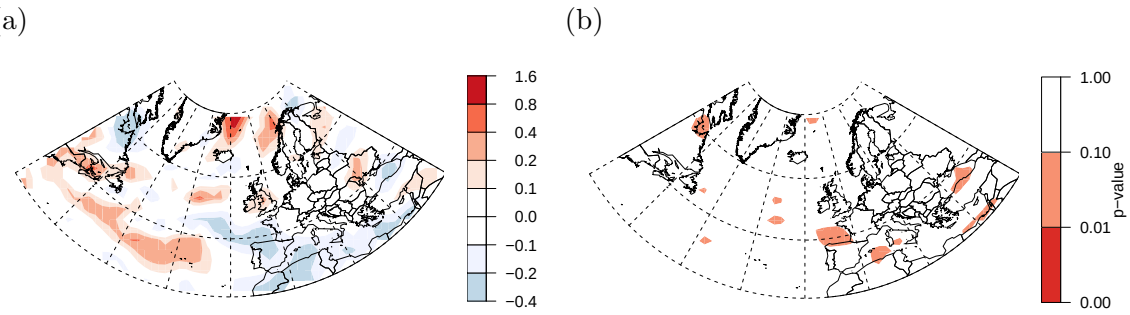

Figure 5.5.: (a) The posterior mean of the emergent constraint λ, and (b) the approximated p-value of the t test for non-zero emergent constraint, estimated from the exchangeable ensemble defined in Chapter 4.

In Chapter 4, FGOALS-g2, MIROC-ESM and MIROC-ESM-CHEM were all poorly predicted in the marginal cross-validation of the model mean climate response in the full ensemble. The future mean climates of MIROC-ESM and MIROC-ESM-CHEM are also poorly predicted in the conditional cross-validation (Figure 5.6). A closer inspection reveals that all three models are outlying in the subtropics and are influencing the estimation of the emergent relationship (Figure 5.4b). Removing any one model from the full ensemble is not sufficient to disrupt the estimated relationship. If all three models are removed, then no significant evidence of an emergent relationship remains when estimated by ensemble regression. This example also demonstrates the dangers of estimating emergent constraints from a small number of models. The extended hierarchical framework actually estimates a positive emergent relationship in the subtropics from the thinned ensemble (Figure 5.4b).

In the full ensemble, the estimated emergent relationship was negative over most of the study area (Figure 5.2). When the ensemble is thinned, the estimated relationship becomes patchy and is only significantly different from zero at a handful of scattered grid boxes (Figure 5.5b). It is possible that the three models identified as influential in the subtropics are in fact contributing valuable information about some underlying physical relationship. However, all three belong to the group identified as having extreme biases in the historical scenario. They also all have relatively low resolution in the atmosphere, and so are unlikely to accurately simulate the structure of the storm track (Colle et al., 2013). Therefore, we conclude that there is no robust evidence of an emergent constraint on cyclone frequency in the North Atlantic.

5.6. Discussion

The hierarchical framework of Chapter 4 has been extended to incorporate correlation between the expected climate responses and historical climates of the models.

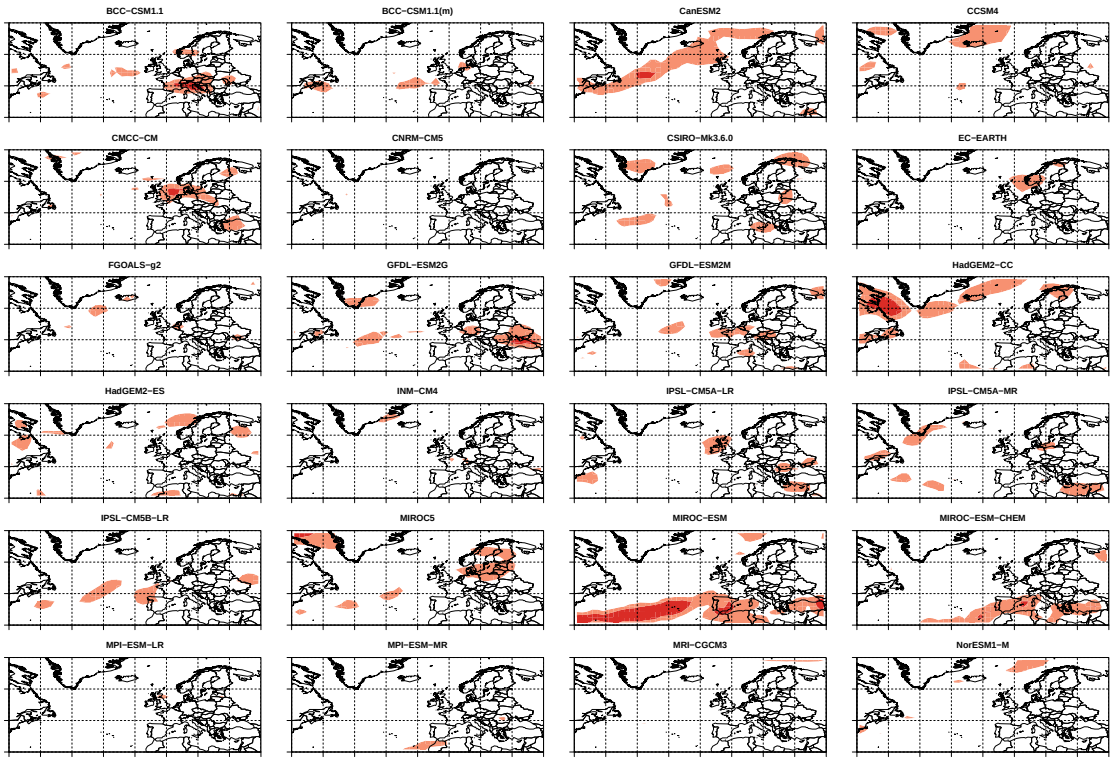

Figure 5.6.: The p-values from the conditional cross-validation on the model mean future climates, shading is the same as Figure 5.5b.

It is usually assumed that the historical and future departures of a particular run due to internal variability (weather) are independent, over long time scales. The extended hierarchical framework explicitly incorporates this assumption by estimating the correlation at the level of the random effects representing the model departures. Other studies incorporating correlations between model outputs have included only one run from each model (e.g., Tebaldi et al., 2005), or the means of the runs from each model (e.g., Bracegirdle and Stephenson, 2012). In either case, model differences are confounded with internal variability. This can actually lead to a negative bias in the estimation of any emergent relationship. The extended hierarchical framework derived here separates model differences from internal variability and so is not susceptible to this bias. Once again, this highlights the importance of accounting for internal variability when analysing the outputs of an ensemble of climate models.

Emergent relationships have previously been exploited to constrain projections of the actual climate (Hall and Qu, 2006; Boé et al., 2009). This assumes that the relationship between the models corresponds to some real physical process in the actual climate. If the relationship is simply an artefact of the models, then the projection will be biased rather than corrected by applying the emergent constraint (Knutti et al., 2010b). Therefore, unless the underlying physical process is understood, projections with and without the emergent constraint should be presented

and contrasted. Emergent relationships have also been used to constrain our uncertainty about future climate (Bracegirdle and Stephenson, 2012; Cox et al., 2013; Karpechko et al., 2013). The uncertainty about the actual climate response is assumed to be the same as the uncertainty about the response of a new climate model, conditional on its historical climate. This implies an assumption of exchangeability between the models and the actual climate. As discussed in Chapter 2, this judgement may be too strong. A more flexible alternative is proposed in Chapter 6.

The marginal cross-validation approach to framework checking described by Smith et al. (2009) has been extended here to check the assumptions about the conditional distribution of the model responses when an emergent constraint is estimated. The example of the North Atlantic storm track illustrated the importance of checking for influential models. It also demonstrated that standard leave-one-out cross-validation may not be sufficient. In principle, the cross-validation approach is easily extended to leaving out all possible pairs, triples, etc. of models. This may be feasible for a single study area or where data are aggregated into a small number of study areas. However, fitting the hierarchical framework is computationally intensive and so this may not be possible for a large number of regions / grid boxes. One possibility would be to use ensemble regression and the Cook's distance diagnostics suggested by Bracegirdle and Stephenson (2012) to approximate the conditional cross-validation.

The ensemble regression approach of Bracegirdle and Stephenson (2012) only considers correlations between the response of a particular climate variable and the historical state of the same variable. Other studies have correlated the response of one variable with the historical state of another (e.g., Boé et al., 2009; Hall and Qu, 2006). Ensemble regression has also been extended to a multiple regression approach, i.e., the correlation between one response variable and multiple explanatory historical variables (Karpechko et al., 2013). Single or multiple covariates could also be included in the hierarchical framework. However, the results in this chapter highlight the importance of accounting for internal variability in order to avoid biasing estimates of the correlations. Therefore additional variables should be treated not simply as known covariates but as additional response variables with their own uncertainties due to model differences and internal variability. The hierarchical framework described here can be easily extended to the multivariate case. The derivation of such a framework is beyond the focus of this thesis, however.

Checking the fit of the extended hierarchical framework to the extra-tropical cyclone data demonstrated the sensitivity of emergent relationships to outlying models. The strong correlation between the climate responses and historical climates of the models noted in the subtropics in Chapter 3 appears to be the product of three poorly performing models. The models in question had already been excluded from the exchangeable ensemble due to extreme biases and low resolution compared to the

rest of the ensemble. So thinning the ensemble had a positive effect by removing their influence. However, great care must be taken when making judgements about the exchangeability of model outputs. Excluding models unnecessarily will lead to a loss of information and decrease the precision of the parameter estimates, making them more susceptible to influence by outlying models.

In Chapter 2 we speculated about the existence of possible emergent constraints on the response of extra-tropical cyclone frequency to climate change. From the analysis here, we conclude that there is no robust evidence of an emergent constraint on cyclone track density in the North Atlantic, at least at the grid box level. Therefore, projections of track density should be based on a framework that does *not* include an emergent constraint, in order to avoid spurious biases. This does not preclude the possibility of constraining projections of cyclone frequency. Other studies have found evidence of emergent constraints when climate model output is aggregated over large spatial scales (Chang et al., 2012, 2013),or when cyclone frequency is linked to other variables (Woollings et al., 2012; Harvey et al., 2013). Investigating such relationships is beyond the scope of this thesis, however.

5.7. Conclusions

In this chapter, the hierarchical framework developed in Chapter 4 was extended to incorporate correlations between the climate responses and historical climates of the models. The resulting framework improves upon established simple linear regression approaches by correcting for a bias in the estimation of the emergent relationships. It also improves upon earlier Bayesian approaches by separating model differences from departures due to internal variability. The formulation of the framework derived here explicitly reflects the assumption that emergent constraints should apply only to differences between models and not to differences between runs from the same model.

Emergent constraints have the potential to provide valuable information about the future state of the actual climate. Applying emergent relationships to constrain future projections in the manner of ensemble regression requires strong assumptions about the relationship between the ensemble and the actual climate. Specifically, it requires the assumption that the actual climate is exchangeable with the models. The relationship between the ensemble and the actual climate is considered in detail in Chapter 6, where a less less restrictive assumption is proposed.

6. How to relate multi-model ensembles to the actual climate

6.1. Introduction

The hierarchical frameworks derived in the previous chapters allow us to make inferences about the outcomes of new model runs, or the expected climates of new models, but not about the climate of the Earth system itself. Two main interpretations of the relationship between climate models and the Earth system exist. The "truth plus error" paradigm treats climate model output as the actual climate (the expected value of the distribution of weather in the Earth system, Section 2.1) plus some error. Alternatively, the "exchangeable" paradigm treats the Earth system as though it were just another model, so that the actual climate and the expected climate of each model are drawn from the same underlying distribution. In Chapter 2, it was argued that neither approach was wholly satisfactory.

Chandler (2013) and Rougier et al. (2013) proposed the idea of including a discrepancy between the expected climate of an ensemble of climate models and the actual climate. The method proposed by Chandler (2013) is a generalisation of the "truth plus error" approach, while the method proposed by Rougier et al. (2013) is a generalisation of the "exchangeable" approach.

This chapter extends the approach proposed by Rougier et al. (2013) to include the effects of uncertainty due to internal variability between model runs, and emergent constraints between the climate responses and historical climates of the models. Emergent relationships are reinterpreted as providing constraints on the discrepancy between the expected response of the ensemble and the actual climate response. Further extensions are proposed in order to allow sampling uncertainty about the actual climate due to natural variability in the Earth system, and measurement error in the observations to be treated separately.

In addition, it is shown that identical inferences will be obtained by adopting either of the general approaches proposed by Rougier et al. (2013) and Chandler (2013), if identical assumptions are made about key components. The importance of account-

ing for internal variability when estimating emergent constraints was demonstrated in Chapter 5. In this chapter, it is shown that sampling uncertainty and measurement error play an important role when making projections based on emergent relationships.

6.2. The ensemble and the actual climate

The standard approach for linking computer models such as climate models to physical systems is to relate the expected value of the model to the expected value of the system (Craig et al., 2001; Kennedy and O'Hagan, 2001). It is assumed that the model is run at its best parameter settings, so that nothing further can be learned about the state of the real system from adjusting the inputs to the model (Craig et al., 2001). This is compatible with the interpretation of a multi-model ensemble as a collection of "best guesses". No modelling group wants its model to be seen to perform poorly, so the "best" known configuration is likely to be submitted to the ensemble. Rougier et al. (2013) and Chandler (2013) generalised this approach to an ensemble of models by relating the expected climate of the ensemble to the expected climate of the Earth system. The expected climate of the ensemble is interpreted as our "best" estimate of the actual climate given our theoretical knowledge of the Earth system and our ability to model it.

6.2.1. The expectation of the historical climate

We begin by specifying the relationship between the ensemble of climate models and the actual climate in the historical period as

$$y_H = \mu + \Delta_H \tag{6.1a}$$

$$\Delta_H \sim N\left(0\ ,\ \sigma^2_{\Delta_H}\right) \tag{6.1b}$$

where y_H is the actual historical climate (the expectation of the historical distribution of weather), and μ is the expected climate of the ensemble. It can be helpful to interpret μ as the expected climate of a representative model (Rougier et al., 2013). The Δ_H term represents the discrepancy between the expected climate of the models and the actual climate. The discrepancy reflects the fact that all climate models are imperfect representations of the climate system. It is treated as a random quantity with associated variance $\sigma^2_{\Delta_H}$, which quantifies our beliefs about how informative the models are for the actual climate.

6.2.2. The expectation of the future climate

We take a similar approach in order to represent the relationship between the ensemble and the actual future climate. We relate the expected future climate of the ensemble to the actual climate in the future period by

$$y_F = \mu + \Delta_H + \beta + \Delta_R \tag{6.2a}$$

$$\Delta_R \mid \Delta_H \sim N\left(\lambda\Delta_H \ , \ \sigma^2_{\Delta_R|\Delta_H}\right) \tag{6.2b}$$

where y_F is the actual future climate (the expectation of the distribution of future weather), β is the expected climate response of the ensemble, and λ is the emergent constraint estimated from the ensemble. The Δ_R term represents the discrepancy between the expected response of the models and the actual climate response. Like the historical discrepancy Δ_H, the response discrepancy is treated as a random quantity. However, it is conditioned on the historical discrepancy in a manner analogous to the response departures of the climate models in the hierarchical framework proposed in Chapter 5. Therefore, we obtain an ensemble regression-like linear adjustment to the actual climate response y_R (the change in the expectation of the distribution of weather)

$$\mathrm{E}\left(y_R\right) = \mathrm{E}\left(y_F - y_H\right) = \beta + \lambda\Delta_H \tag{6.3}$$

The expectation of the actual climate response depends on the discrepancy Δ_H between the actual historical climate and the expected historical climate of the ensemble μ, in the same way that the expected response of model m depends on its historical departure (α_m) from μ, i.e., the emergent relationship λ is assumed to represent a physical constraint that will apply equally to the actual climate. The variance $\sigma^2_{\Delta_R|\Delta_H}$ quantifies our beliefs about how informative the models are for the actual climate response. From Equations 6.2b and 6.1b, the marginal uncertainty about actual climate response y_R is

$$\sigma^2_{\Delta_R} = \lambda^2\sigma^2_{\Delta_H} + \sigma^2_{\Delta_R|\Delta_H}$$

Clearly $\sigma^2_{\Delta_R|\Delta_H} < \sigma^2_{\Delta_R}$ for $\lambda \neq 0$, and so emergent relationships reduce our uncertainty about the actual climate response, given knowledge of the historical state.

6.2.3. Sampling uncertainty and natural variability

In Chapter 2, climate was defined as the distribution of weather, and the actual climate as the expectation of that distribution. Operationally, climate is defined as the 30-year average of weather. The weather we experience over any 30-year period

is one sample or *actualisation* (as opposed to a *realisation* of a model) from the distribution of possible weather. Therefore, as an estimate of the actual climate, it is subject to sampling uncertainty equivalent to the variance of the historical distribution of 30-year weather (the natural variability, Chapter 2). Similarly, the future climate that we will experience is only on sample from the distribution of possible future weather and is subject to natural variability equivalent to the variance of the future distribution of 30-year weather. Let y_{Ha} be the historical climate (average of weather) that we have experienced, and y_{Fa} be the future climate that we might experience, then

$$y_{Ha} = N\left(y_H \ , \ \sigma^2_{Ha}\right) \tag{6.4a}$$

$$y_{Fa} = N\left(y_F \ , \ \sigma^2_{Fa}\right) \tag{6.4b}$$

where the variances σ^2_{Ha} and σ^2_{Fa} represent the sampling uncertainty associated with the historical climate and the uncertainty due to natural variability in the future climate, respectively. The sampling uncertainty in the historical scenario σ^2_{Ha} can be estimated from the time series of observations within the historical period. However, the natural variability in the future scenario σ^2_{Fa} must either be assumed to be constant and equal to that of the historical scenario, or estimated from the models. One possible approach would be to assume that

$$\sigma^2_{Fa} = \theta\sigma^2_{Ha} \quad \text{where} \quad \theta = \frac{\sigma^2_F}{\sigma^2_H} \tag{6.5}$$

similar to Tebaldi et al. (2005), i.e., the fractional change in variability in the weather of the Earth system is equal to that simulated by the models. If the climate models were to simulate the variability of the weather perfectly, then $\sigma^2_{Ha} = \sigma^2_H$ and $\sigma^2_{Fa} = \sigma^2_F$. Therefore, an estimate of σ^2_{Ha} based on climate model output can be substituted if necessary (e.g., Collins et al., 2013, Box 12.1).

6.2.4. Observation uncertainty

In addition to sampling uncertainty, our knowledge of the actual historical climate y_H is limited by our ability to measure it accurately. Our observations will differ from the climate we actually experience y_{Ha} for a variety of reasons, e.g., poorly positioned weather stations, instrument biases, human error. These errors are usually assumed to be small and so are often neglected (Gleckler et al., 2008). However, the quality of observation data varies considerably (Thorne et al., 2011). Where observations are sparse, the uncertainty may be considerable. Therefore, we explicitly include

the effect of observation uncertainty as

$$z = N\left(y_{Ha}\ ,\ \sigma_z^2\right) \tag{6.6}$$

where z are the observations, and the variance σ_z^2 is an estimate of the uncertainty due to observation errors, i.e., it quantifies how informative we believe our observations are for the climate we experience. We only consider observation uncertainty for the historical climate. It is problematic to estimate observation uncertainty for the future climate. By the time the future climate is observed, the observing network and methods will have changed dramatically.

6.2.5. The complete framework

The proposed framework in its entirety, including the hierarchical framework developed in Chapter 5 can be summarised as

$$\begin{aligned}
x_{Hmr} &\overset{iid}{\sim} N\left(\mu + \alpha_m\ ,\ \sigma_H^2\right) & y_{Ha} &\sim N\left(\mu + \Delta_H\ ,\ \sigma_{Ha}^2\right)\\
x_{Fmr} &\overset{iid}{\sim} N\left(\mu + \alpha_m + \beta + \gamma_m\ ,\ \sigma_F^2\right) & y_{Fa} &\sim N\left(\mu + \Delta_H + \beta + \Delta_R\ ,\ \sigma_{Fa}^2\right)\\
\alpha_m &\overset{iid}{\sim} N\left(0\ ,\ \sigma_\alpha^2\right) & \Delta_H &\sim N\left(0\ ,\ \sigma_{\Delta_H}^2\right)\\
\gamma_m \mid \alpha_m &\overset{iid}{\sim} N\left(\lambda\alpha_m\ ,\ \sigma_{\gamma|\alpha}^2\right) & \Delta_R \mid \Delta_H &\sim N\left(\lambda\Delta_H\ ,\ \sigma_{\Delta_R|\Delta_H}^2\right)
\end{aligned}$$

$$z \sim N\left(y_{Ha}\ ,\ \sigma_z^2\right) \tag{6.7}$$

The basic framework without an emergent constraint is a special case where $\lambda = 0$. The complete framework is illustrated as a directed acyclic graph in Figure 6.1. The graph is a useful aid to understanding the dependencies between the various components when making inferences about the actual climate later in the chapter.

6.3. Making judgements about the ensemble discrepancies

The ensemble discrepancies Δ_H and Δ_R arise due to the fact that all climate models are imperfect representations of the climate system. The variances $\sigma_{\Delta_H}^2$ and $\sigma_{\Delta_R|\Delta_H}^2$ associated with the ensemble discrepancies quantify how informative we believe the climate models are for the actual climate and climate response. Climate models share common limitations such as finite resolution, and some physical processes are not represented in any model. This might lead us to believe that our uncertainty about the actual climate response is greater than our uncertainty about the response

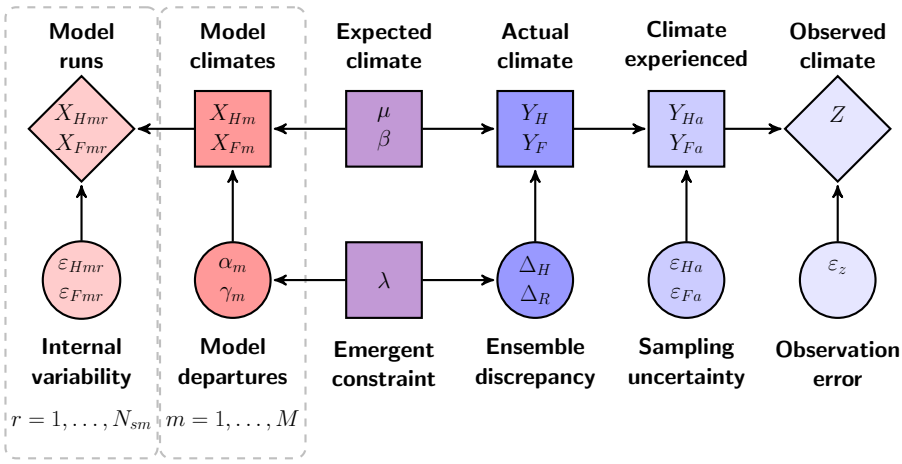

Figure 6.1.: The full framework relating the ensemble to the actual climate and the observations represented as a directed acyclic graph. Diamonds indicate observed or measured quantities, squares indicate latent (unobservable) quantities, and circles indicate mean zero random departures. Arrows indicate the direction of conditioning.

a new model. In the absence of any other data (e.g., observations), a similar argument can be made for our uncertainty about the actual historical climate. In other words

$$\sigma_\alpha^2 \leq \sigma_{\Delta_H}^2 \quad \text{and} \quad \sigma_{\gamma|\alpha}^2 \leq \sigma_{\Delta_R|\Delta_H}^2$$

which is equivalent to the third of the simple criteria for a credible representation of the ensemble and the actual climate identified in Chapter 2, i.e., our uncertainty about the actual climate response should span the spread of responses simulated by the climate models. Following Rougier et al. (2013), such a judgement may be represented by a scaling factor, so that

$$\sigma_{\Delta_H}^2 = \kappa^2 \sigma_\alpha^2 \quad \text{and} \quad \sigma_{\Delta_R|\Delta_H}^2 = \kappa^2 \sigma_{\gamma|\alpha}^2 \tag{6.8}$$

for some constant $\kappa \geq 1$. This should not be interpreted as implying any relationship between the discrepancies and the model departures. It is simply a device that allows us to express our prior beliefs about the ensemble discrepancies, relative to the model spread. Applying the same scaling factor to both the historical and response uncertainty expresses the natural judgement that the models are no more informative for the climate response than they are for the historical climate (relative to the spread in the models). In other words, why would we believe that the models can simulate the climate response correctly, if they cannot reproduce the basic state of the system to a reasonable approximation? In the absence of strong beliefs, $\kappa = 1$ might be regarded as the default choice. This is equivalent to assuming that the expected climates of the models are exchangeable with (or "statistically indistinguishable" from) the actual climate (Rougier et al., 2013). This judgement may be too strong, as argued above and in Chapter 2. However, this is likely to be

the most common judgement as climate scientists begin to allow for discrepancies between climate models and the actual climate.

6.4. Combining model outputs with observations

The framework described by Equation 6.7 combines information from climate model outputs and from observations of the actual climate. Our beliefs about the climate of the Earth system given both the model outputs *and* the observations are quantified by the posterior distribution

$$\Pr\left(\mathbf{y} \mid \mathbf{x}, z\right)$$

where $\mathbf{y} = (y_H, y_{Ha}, y_F, y_{Fa})$ is the vector of random quantities describing the climate of the Earth system, $\mathbf{x}$ are the climate model outputs, and z is the observations. In Appendix D.1, it is shown that if the marginal posterior distribution of the expected climate of the ensemble $\Pr\left(\mu \mid \mathbf{x}\right)$ is known (e.g., from the hierarchical framework developed in Chapter 5) and well approximated by

$$\mu \mid \mathbf{x} \sim N\left(\mu_\mu, \sigma_\mu^2\right)$$

then the posterior distribution of the actual historical climate y_H is

$$\Pr\left(y_H \mid \mathbf{x}, z\right) \sim N\left(\frac{\tau_{y|x}\mu_\mu + \tau_{z|y}z}{\tau_{y|x} + \tau_{z|y}} \;,\; \left(\tau_{y|x} + \tau_{z|y}\right)^{-1}\right) \tag{6.9}$$

where

$$\tau_{z|y} = \left(\sigma_{Ha}^2 + \sigma_z^2\right)^{-1} \quad \text{and} \quad \tau_{y|x} = \left(\sigma_\mu^2 + \sigma_{\Delta_H}^2\right)^{-1}$$

Therefore, the expectation of the posterior distribution of the actual historical climate y_H is a weighted average of the expected climate of the ensemble μ and the observed climate z. The posterior expectation of the difference between the actual historical climate y_H and the observations z is

$$\mathrm{E}\left(y_H - z \mid \mathbf{x}, z\right) = \frac{\tau_{y|x}\left(\mu_\mu - z\right)}{\tau_{y|x} + \tau_{z|y}}$$

So the posterior estimate of the actual historical climate experiences a shrinkage towards the expected climate of the models that depends on the variance ratio

$$I = \frac{\tau_{y|x}}{\tau_{y|x} + \tau_{z|y}} \tag{6.10}$$

which we will call the information ratio. The information ratio I is bounded between 0 and 1. The greater the information ratio, the greater the shrinkage of the posterior estimate of the actual climate towards the expected climate of the models and away

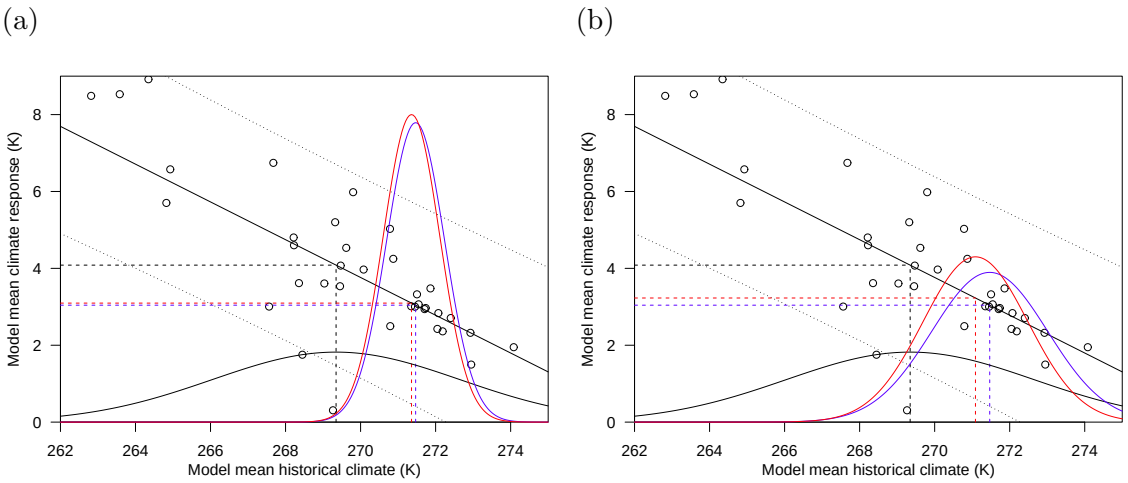

Figure 6.2.: Examples of projection using emergent constraints where (a) the models are uninformative compared to the observations ($I \approx 0.05$); and (b) the models are mildly informative compared to the observations ($I \approx 0.18$). The solid black line represents the emergent relationship between the historical climate and climate response. The black dotted lines are a 95% prediction interval for the response of a new model based on ensemble regression. The dashed lines represent the mean historical climate and climate response according to the posterior distribution given the models (black), the observations (blue), and the posterior distribution given the models and the observations (red).

from the observations. Therefore, it is not simply how informative the models are judged to be for the actual climate that matters, but how informative they are compared to the observations given the estimated measurement error and sampling uncertainty.

An interesting question is what effect does the adjustment of our beliefs about the actual historical climate have on projections? From Equation 6.3, the expectation of the actual climate response is

$$\mathrm{E}\left(y_F - y_H\right) = \beta + \lambda \Delta_H$$

When there is no emergent relationship (i.e, $\lambda = 0$), then the expected response of the actual climate is equal to the expected response of the ensemble β. However, if there is an emergent relationship, then the projected response depends on the historical discrepancy Δ_H. The posterior expectation of the historical discrepancy Δ_H is

$$\mathrm{E}\left(\Delta_H \mid \mathbf{x}, z\right) = \mathrm{E}\left(y_H - \mu \mid \mathbf{x}, z\right) = \frac{\tau_{z|y}\left(z - \mu_\mu\right)}{\tau_{y|x} + \tau_{z|y}} = \left(1 - I\right)\left(z - \mu_\mu\right)$$

If the models are not informative compared to the observations ($I \approx 0$), then the posterior expectation of the historical discrepancy is simply the difference between the observations and the expected climate of the ensemble ($\mathrm{E}\left(\Delta_H\right) \approx z - \mu_\mu$). If

the historical discrepancy is large, then the projected climate response may be quite different to the expected response of the ensemble (Figure 6.2a). On the other hand, if the models are judged to be relatively informative compared to the observations ($I > 0$), then the historical discrepancy Δ_H will decrease as the actual climate experiences shrinkage away from the observed climate z, and towards the expected climate of the models μ. In that case, the difference between the projection of the actual climate response and the expected response of the models will also be reduced (Figure 6.2b). So both observation uncertainty and natural variability play an important role in the projection of future climate, particularly when an emergent relationship is estimated.

6.5. Comparison with previously published methods

In this section, we briefly compare the framework derived in this chapter with previously published statistical frameworks for relating ensembles of climate models to the Earth system. In Chapter 2, three simple criteria for a credible representation of the relationship between the ensemble and the Earth system were identified from the literature. A credible framework should predict that the model biases compared to the observed climate are correlated, that the mean squared error of the multi-model mean should not converge to zero with increasing ensemble size, and that the uncertainty about the actual climate response should be at least as great as that about the response of a new model. The framework proposed here satisfies all three criteria

$$\begin{aligned}
&\mathrm{cov}\left(\bar{x}_{Hir} - y_H \ , \ \bar{x}_{Hjr'} - y_H\right) = \sigma^2_{\Delta_H} \\
&\mathrm{E}\left(\left(\frac{1}{M}\sum_{m=1}^{M}\bar{x}_{Hm.} - y_H\right)^2\right) = \sigma^2_{\Delta_H} + \frac{1}{M}\sigma^2_{\alpha} + \frac{1}{M^2}\sum_{m=1}^{M}\frac{1}{{N_{Hm}}^2}\sigma^2_H \\
&\mathrm{var}\left(y_R\right) = \mathrm{var}\left(\beta\right) + \sigma^2_{\Delta_R} = \mathrm{var}\left(\beta\right) + \kappa^2\sigma^2_{\gamma} \quad (\kappa \geq 1)
\end{aligned}$$

6.5.1. Frameworks including discrepancy terms

Among the published methods reviewed in Chapter 2, only the frameworks proposed by Chandler (2013) and Rougier et al. (2013) were able to fully satisfy our three simple credibility criteria. The two frameworks take very similar approaches and are compared graphically in Figure 6.3. Both frameworks include the idea of a discrepancy between the expected climate of an ensemble of climate models and the actual climate, however they differ in their conditioning assumptions. Rougier

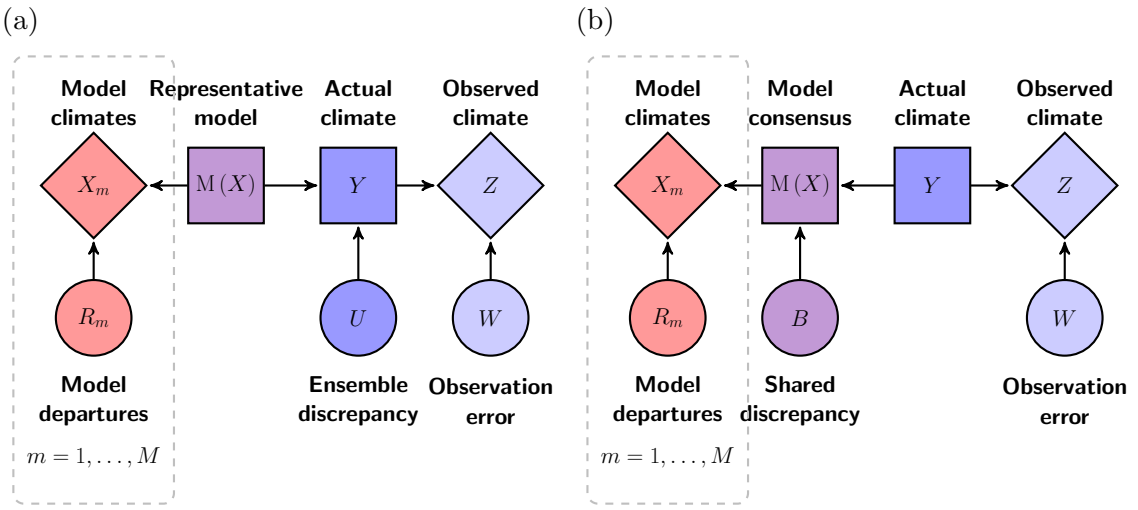

Figure 6.3.: The frameworks proposed by (a) Rougier et al. (2013), and (b) Chandler (2013) illustrated as directed acyclic graphs. Diamonds indicate observed or measured quantities, squares indicate latent (unobservable) quantities, and circles indicate mean zero random departures. Arrows indicate the direction of conditioning. The component representing internal variability in the climate models included in the framework proposed by Chandler (2013) is neglected to simplify the comparison.

et al. (2013) condition the actual climate on the expected climate of the models ($Y = \mathrm{M}(X) + U$ in Figure 6.3a), while Chandler (2013) takes the opposite approach ($\mathrm{M}(X) = Y + B$ in Figure 6.3b). It can be shown that both frameworks will yield identical inferences for the actual climate Y, provided that identical distributional assumptions are made for key components (see Appendix D.2). Therefore, why choose one formulation over the other? As discussed in Chapter 2, the approach taken by Chandler (2013) is a generalisation of the "truth plus error" interpretation of the relationship between climate models and the actual climate. Although the "truth plus error" approach is widely used in climate science, the alternative approach proposed by Rougier et al. (2013) is arguably the more natural formulation. Statistical frameworks are usually constructed for a quantity of interest (e.g., the actual climate) based on some explanatory data (e.g., climate model outputs). The "truth plus error" interpretation takes the opposite approach. This peculiarity was noted by Rougier et al. (2013), who remarked in an earlier draft of their manuscript that "It is doubtful that anyone using [the "truth plus error" approach] feels the need to perform a probabilistic inversion to update their judgements about [the actual climate]". Berliner and Kim (2008) concluded that the direction of conditioning was not important, and should be decided by our ability to formulate the relevant distributions, to interpret them and to perform the necessary computations.

In general, the standard statistical modelling approach of specifying beliefs about a quantity of interest based on explanatory data is easier to interpret. The framework developed in this thesis can be be viewed as an extension of the framework proposed

by Rougier et al. (2013), as can be seen by comparing Figure 6.3 with Figure 6.1. Compatibility with the framework of Rougier et al. (2013) in Equation 2.14 is easily established by writing

$$\mathbf{X}_m = \begin{pmatrix} X_{Hm} \\ X_{Fm} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_H \\ Y_F \end{pmatrix} \quad \text{and} \quad \mathbf{M}(\mathbf{X}) = \begin{pmatrix} \mu \\ \mu + \beta \end{pmatrix} \tag{6.11}$$

where $X_{Hm} = \mu + \alpha_m$ and $X_{Fm} = \mu + \alpha_m + \beta + \gamma_m$ are the expected historical and future climates of model m. Then the model departures and the structural discrepancy can be re-expressed in terms of multi-variate normal distributions so that

$$\mathbf{R}_m = \begin{pmatrix} \alpha_m \\ \gamma'_m \end{pmatrix} \overset{iid}{\sim} MVN\left(\mathbf{0}\ ,\ \begin{pmatrix} \sigma^2_\alpha & (1+\lambda)\,\sigma^2_\alpha \\ (1+\lambda)\,\sigma^2_\alpha & (1+\lambda)^2\,\sigma^2_\alpha + \sigma^2_{\gamma|\alpha} \end{pmatrix}\right) \tag{6.12}$$

where γ'_m is the departure of model m from the expected future climate of the ensemble (rather than the expected climate response), and

$$\mathbf{U} = \begin{pmatrix} \Delta_H \\ \Delta_F \end{pmatrix} \sim MVN\left(\mathbf{0}\ ,\ \begin{pmatrix} \sigma^2_{\Delta_H} & (1+\lambda)\,\sigma^2_{\Delta_H} \\ (1+\lambda)\,\sigma^2_{\Delta_H} & (1+\lambda)^2\,\sigma^2_{\Delta_H} + \sigma^2_{\Delta_R|\Delta_H} \end{pmatrix}\right) \tag{6.13}$$

where Δ_F is the discrepancy between the actual future climate and the expected future climate of the ensemble. The hierarchical framework developed in Chapter 5 extends the framework of Rougier et al. (2013) by allowing for uncertainty in the expected climates of the models due to internal variability, i.e., we model X_{Hmr} and X_{Fmr} in addition to X_{Hm} and X_{Fm}. The framework proposed here also extends the formulation of Rougier et al. (2013) by allowing the separation of the effects of sampling uncertainty about the actual climate, and measurement error in the observations.

6.5.2. Methods including emergent constraints

Of the published methodologies reviewed in Chapter 2, only two explicitly included the estimation of an emergent constraint. The ensemble regression approach proposed by Bracegirdle and Stephenson (2012) has already been discussed in Chapters 2 and 5. A heuristic interpretation of the framework in graphical form is given in Figure 6.4a. The projected response y_R in ensemble regression has the same basic form as the framework developed here (Equation 6.3)

$$\mathrm{E}\,(y_R) = \beta + \lambda\,(z - \bar{x}_{H.})$$

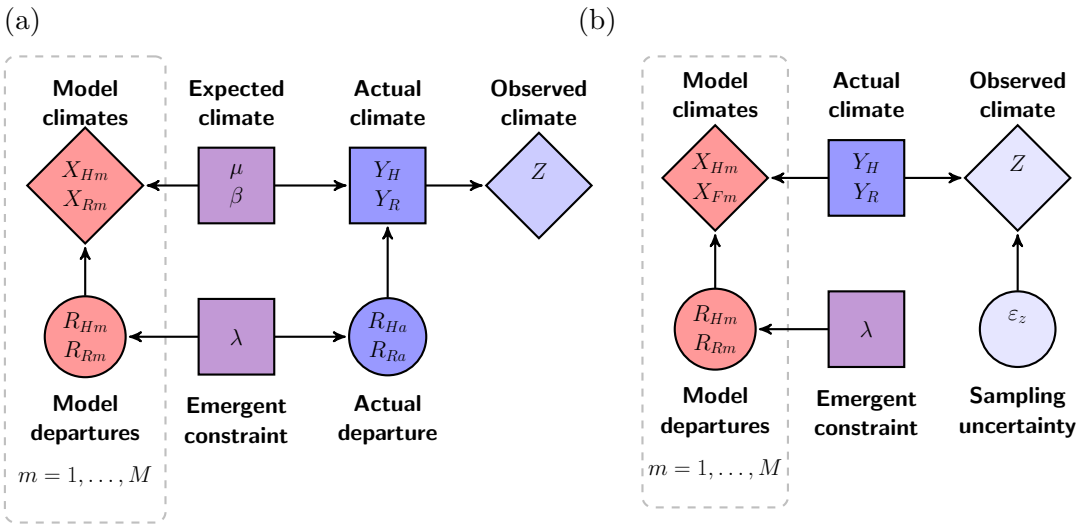

Figure 6.4.: The frameworks proposed by (a) Bracegirdle and Stephenson (2012), and (b) Tebaldi et al. (2005) (alternative formulation using Equation 6.15) illustrated as directed acyclic graphs. Diamonds indicate observed or measured quantities, squares indicate latent (unobservable) quantities, and circles indicate mean zero random departures. Arrows indicate the direction of conditioning. Note that Bracegirdle and Stephenson (2012) treat all historical quantities as fixed rather than random quantities, and assume that $\mathrm{var}\,(R_{Rm}) = \mathrm{var}\,(R_{Ra})$. As noted in the text, the Normal-Gamma mixture formulation of Tebaldi et al. (2005) is equivalent to treating the model outputs as a random sample from a t distribution, so the τ_m terms are neglected in this comparison.

Therefore, we expect the estimates of the expected value of the projected response to be similar. In Chapter 5, it was noted that Bracegirdle and Stephenson (2012) implicitly assume that the actual climate response is exchangeable with the expected responses of the models. So under the default assumption of $\kappa = 1$ in the framework developed here, the uncertainty associated with the projected response will also be similar. In general, the posterior uncertainty from the framework derived here is expected to be slightly greater since ensemble regression does not account for uncertainty due to internal variability, sampling uncertainty or observation error. The estimates from the two frameworks will diverge if either the shrinkage described in Section 6.4, or the bias in the estimate of the emergent constraint by ensemble regression is large.

The framework proposed by Tebaldi et al. (2005) included a slightly different definition of an emergent constraint. The univariate extension proposed by Smith et al. (2009) can be written in the notation of this thesis as

$$x_{Hm} \overset{iid}{\sim} N\left(y_H \ , \ \tau_m^{-1}\right) \tag{6.14a}$$

$$x_{Fm} \mid x_{Hm} \overset{iid}{\sim} N\left(y_F + \lambda'\left(x_{Hm} - y_H\right) \ , \ \left(\theta\tau_m\right)^{-1}\right) \tag{6.14b}$$

$$\tau_m \overset{iid}{\sim} Gamma\left(k \ , \ l\right) \tag{6.14c}$$

$$z \sim N\left(y_H \ , \ \tau_z^{-1}\right) \tag{6.14d}$$

where λ' represents a correlation between the future climate of model m and its historical climate, rather than the climate response of model m and its historical climate. The τ_m are model specific precisions that are assumed to arise from a common Gamma distribution with unknown shape and rate parameters k and l. In the original formulation proposed by Tebaldi et al. (2005) the parameters of the Gamma distribution were fixed *a priori*. The parameter θ is a scaling factor that allows the precision of the models to differ for the future climate compared to the historical climate.

It is convenient to rewrite Equation 6.14b in terms of the climate response y_R and the usual emergent constraint λ. Let $y_F = y_H + y_R$ and $\lambda' = \lambda + 1$, then

$$x_{Fm} \mid x_{Hm} \overset{iid}{\sim} N\left(y_H + (x_{Hm} - y_H) + y_R + \lambda\,(x_{Hm} - y_H)\ ,\ (\theta\tau_m)^{-1}\right) \tag{6.15}$$

so the future climate of model m is made up of the actual historical climate y_H, a departure from the historical climate $x_{Hm} - y_H$, the actual climate response y_R, and a departure from the actual response that is proportional to the historical departure $\lambda\,(x_{Hm} - y_H)$. The basic structure is similar to the hierarchical framework developed in Chapter 5, except that the models are centred on the actual climate y_H and climate response y_R, rather than the expected values of the ensemble μ and β. The alternative parameterisation using Equation 6.14b is shown in graphical form in Figure 6.4b for comparison with the other frameworks.

The Normal-Gamma mixture formulation used by Smith et al. (2009) is equivalent to assuming that the model departures from the actual climate are a random sample from a t distribution (Gelman et al., 2014, page 437). The posterior means of the actual historical climate y_H and climate response y_R in this formulation are weighted averages of the model climates and climate responses, weighted by the model specific precisions τ_m (Tebaldi et al., 2005). In Appendix D.3, it is shown that the posterior expectations of the τ_m themselves are approximately

$$\mathrm{E}\,(\tau_m \mid \ldots) \approx \frac{1}{\frac{(x_{Hm}-y_H)^2}{2} + \frac{\theta(x_{Rm}-y_R-\lambda(x_{Hm}-y_H))^2}{2}} \tag{6.16}$$

where $x_{Rm} = x_{Fm} - x_{Hm}$. So the weight given to each model depends on its departures from the actual historical climate $(x_{Hm} - y_H)$, and from the projected climate response (y_R) allowing for any emergent relationship that is present $(\lambda\,(x_{Hm} - y_H))$. This formulation has been criticised for rewarding models that simulate similar responses (Lopez et al., 2006). However, the t distribution is often used to make inferences robust to the presence of outlying data points (Tebaldi et al., 2005; Gelman et al., 2014). Such robustness is desirable, especially given the small number of models that may be included in an ensemble after the thinning process advocated in the previous chapters.

The posterior expectation of the actual climate response y_R (conditional on the parameters) in the alternative parameterisation of the framework of Tebaldi et al. (2005) using Equation 6.15 can be written as

$$\mathrm{E}\left(y_R \mid \ldots\right) \approx \frac{\sum_{m=1}^{M} \tau_m x_{Rm}}{\sum_{m=1}^{M} \tau_m} + \lambda \left(y_H - \frac{\sum_{m=1}^{M} \tau_m x_{Hm}}{\sum_{m=1}^{M} \tau_m} \right) \tag{6.17}$$

(see Equation D.9 in Appendix D.3). This has a very similar form to the expected response in the framework developed here (Equation 6.3). So assuming that

$$\frac{\sum_{m=1}^{M} \tau_m x_{Hm}}{\sum_{m=1}^{M} \tau_m} \approx \mu \quad \text{and} \quad \frac{\sum_{m=1}^{M} \tau_m x_{Rm}}{\sum_{m=1}^{M} \tau_m} \approx \beta$$

then we might expect the expected responses from the two frameworks to be similar. However, Lopez et al. (2006) showed that the width of the posterior distribution of y_R in the framework proposed by Tebaldi et al. (2005) tends to decrease with the square root of the number of models, and will not span the full spread of climate responses simulated by the models (see also, the conditional variance of y_R in Equation D.9). Therefore, we expect the posterior uncertainty about the actual climate response to be much smaller than in the framework developed here.

In Section 6.4, we showed that the posterior expectation of the actual historical climate y_H in the framework developed here is a weighted average of the observations z and the expected historical climate of the models μ. The posterior expectation of y_H in the framework of Tebaldi et al. (2005) is also a weighted average of the observations and the model outputs. However, Lopez et al. (2006) noted that the models tended to receive more weight than the observations. This is because the posterior uncertainty about y_H tends to be small, similar to y_R above (Equation D.8). So the estimate of the actual historical climate y_H will experience a greater shrinkage towards the (weighted) mean historical climate of the models (Equation 6.9), and the projected climate response y_R will lie closer to the (weighted) mean climate response of the models in Equation 6.17. Therefore, the climate response y_R estimated from the framework of Tebaldi et al. (2005) is expected to lie somewhere between the estimates with and without an emergent constraint by the approach developed here.

6.6. Fitting the full framework

The goal is to find the posterior distribution of the climate of the Earth system given the models and the observations $\Pr\left(\mathbf{y} \mid \mathbf{x}, z\right)$, where $\mathbf{y} = (y_{Fa}, y_F, y_{Ha}, y_H)$. The distribution of interest is obtained by integrating over all other parameters in

the joint posterior

$$\Pr\left(\mathbf{y} \mid \mathbf{x}, z\right) = \int\int \Pr\left(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \tag{6.18}$$

where $\boldsymbol{\theta} = (\alpha_m, \gamma_m \; \forall \; m)$ is the vector of random effects from the hierarchical frameworks, and $\boldsymbol{\phi} = (\mu, \beta, \lambda)$ is the vector of parameters from the hierarchical framework. The variance parameters are neglected for brevity. The directed acyclic graph in Figure 6.1 can be useful for understanding the dependencies in the derivation that follows. The joint posterior of the the actual climate and the parameters can be decomposed by repeated factorisation using the law of conditional probability

$$\begin{aligned}
\Pr\left(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) &= \Pr\left(y_{Fa}, y_F, y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) \\
&= \Pr\left(y_{Fa} \mid y_F, y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}, z\right) \Pr\left(y_F, y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) \\
&= \Pr\left(y_{Fa} \mid y_F\right) \Pr\left(y_F, y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right)
\end{aligned}$$

since y_{Fa} is independent of all other elements given y_F by Equation 6.4b. Then

$$\begin{aligned}
\Pr\left(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) &= \Pr\left(y_{Fa} \mid y_F\right) \Pr\left(y_F \mid y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}, z\right) \Pr\left(y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) \\
&= \Pr\left(y_{Fa} \mid y_F\right) \Pr\left(y_F \mid y_H, \mu, \beta, \lambda\right) \Pr\left(y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right)
\end{aligned}$$

since y_F depends only y_H, μ, β and λ by Equations 6.2 and 6.1a. Factorising again yields

$$\begin{aligned}
\Pr\left(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) &= \Pr\left(y_{Fa} \mid y_F\right) \Pr\left(y_F \mid y_H, \mu, \beta, \lambda\right) \Pr\left(y_{Ha}, y_H, \mid \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}, z\right) \Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) \\
&= \Pr\left(y_{Fa} \mid y_F\right) \Pr\left(y_F \mid y_H, \mu, \beta, \lambda\right) \Pr\left(y_{Ha}, y_H \mid \mu, z\right) \Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right)
\end{aligned} \tag{6.19}$$

since y_{Ha} and y_H are independent of all other elements except μ and z by Equations 6.4a, 6.1a and 6.6, and $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are independent of z. Note that $\Pr\left(y_{Fa} \mid y_F\right)$ is the posterior predictive distribution for y_{Fa}, given by Equation 6.4b. Similarly, $\Pr\left(y_F \mid y_H, \mu, \beta, \lambda\right)$ is the posterior predictive distribution for y_F, given by Equation 6.2. The final term, $\Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right)$, is the posterior distribution of the hierarchical framework developed in Chapter 5.

It only remains to find $\Pr\left(y_{Ha}, y_H \mid \mu, z\right)$, the joint posterior of the actual historical climate and the historical climate that we experienced. The conditional posterior distributions of y_{Ha} and y_H are easily obtained from Equations 6.1, 6.4a and 6.6 as

$$\Pr\left(y_{Ha} \mid y_H, z\right) = N\left(\frac{\tau_{Ha} y_H + \tau_z z}{\tau_{Ha} + \tau_z} \;,\; \left(\tau_{Ha} + \tau_z\right)^{-1}\right) \tag{6.20}$$

and

$$\Pr\left(y_H \mid y_{Ha}, \mu\right) = N\left(\frac{\tau_{Ha} y_{Ha} + \tau_{\Delta_H}\mu}{\tau_{Ha} + \tau_{\Delta_H}} \ , \ \left(\tau_{Ha} + \tau_{\Delta_H}\right)^{-1}\right) \tag{6.21}$$

The factorisation in Equation 6.19 means that if we have already estimated the ensemble parameters using the hierarchical framework, then there is no need to refit it in order to form the joint posterior for the climate of the Earth system. It only remains to make judgements about the prior uncertainty about the ensemble discrepancy, and combine the information from the models with the information from the observations. The fitting procedure can proceed in stages

1. Specify the scaling factor κ;
2. Obtain N samples from the joint posterior of the model parameters (μ, β, λ, σ_H^2, σ_F^2, σ_α^2, $\sigma_{\gamma|\alpha}^2$), as described in Chapter 5;
3. For each of the N samples of the model parameters,
   a) sample one estimate of the actual historical climate y_H from Equation 6.21;
   b) sample one estimate of the historical climate that we experienced y_{Ha} from Equation 6.20;
   c) sample one estimate of the actual future climate y_F from Equation 6.2;
   d) sample one estimate of the future climate that we might experience y_{Fa} from Equation 6.4b;
   e) compute the projected climate response $y_R = y_F - y_H$;
   f) compute the projected climate response that we might experience $y_{Ra} = y_{Fa} - y_{Ha}$.

6.7. Using reanalysis data

Many observation datasets do not include estimates of their associated uncertainty. Examples do exist of carefully constructed estimates of observation uncertainty (e.g., HadCRUT4 Morice et al., 2012), however they are the exception rather than the rule. Given the potential importance of observational uncertainty highlighted in the Section 6.4, how should we proceed when that uncertainty is unknown?

One possibility is to make use of reanalysis data as a proxy for observations. Only one of the current generation of reanalysis products includes any form of estimate of the uncertainty associated with the analysis (The Twentieth Century Reanalysis, Compo et al., 2011). However, we might consider comparing or combining data

from multiple reanalyses. It is important to remember that reanalyses are *not* observations. They are a representation of the state of the Earth system, output from a model after assimilating all available observations. If we were to assume that reanalyses were independent observations of the actual climate, then the observational uncertainty could be reduced to the point of being negligible if enough reanalyses were available. However, reanalyses are based on the same observations, and share similar underlying models and data assimilation techniques. Therefore, it seems more appropriate to allow for a discrepancy between the reanalyses and the actual climate as well. Let v_r represent the output from reanalysis r, then consider the following framework

$$v_r \overset{iid}{\sim} N\left(\nu\ ,\ \sigma_v^2\right) \tag{6.22a}$$

$$y_H = \nu + \Delta_v \tag{6.22b}$$

$$\Delta_v \sim N\left(0, \sigma_{\Delta_v}^2\right) \tag{6.22c}$$

Like the models, we assume that each reanalysis is an equally valid representation of the Earth system. This is equivalent to assuming that the reanalyses are exchangeable with one another. Therefore, they are modelled as a random sample from a normal distribution with expectation ν and variance σ_v^2. The actual climate is modelled as the expectation of the reanalyses plus a discrepancy Δ_V. The discrepancy is assumed to be independent of ν, and of the departures of the individual reanalyses. As before, it is convenient to express our judgements about the prior uncertainty associated with the discrepancy $\sigma_{\Delta_V}^2$ relative to the spread in the reanalyses σ_v^2, so let

$$\sigma_{\Delta_V}^2 = \kappa_v^2 \sigma_v^2 \tag{6.23}$$

Once again, the default assumption might be $\kappa_v = 1$. Assuming that the model underlying the reanalysis performs well, and that observations are plentiful and assimilated at short time intervals, then one might expect that the resulting analysis will closely reflect the observed climate. Even if observations are sparse, then provided that the underlying model simulates all of the relevant processes well, it should still be hoped that the analysis will be a good approximation of the actual climate. However, if observations are sparse and some important processes are not well represented, then the analysis is likely to reflect any discrepancy between the climate of the underlying model and the actual climate. Therefore, it is important to make a realistic assessment of how informative the reanalyses are for the actual climate by careful specification of κ_v or σ_v^2.

The parameters ν and σ_v^2 can be estimated similarly to the parameters of the hierarchical frameworks in Chapters 4 and 5. Vague prior distributions are assumed for

both

$$\nu \sim N\left(a_\nu, b_\nu^{-1}\right) \tag{6.24a}$$

$$\tau_v \sim Gamma\left(c_v, d_v\right) \tag{6.24b}$$

where $\tau_v = \sigma_v^{-2}$. The default choices for the hyper-parameters are $a_\nu = 0$, $b_\nu = 10^{-6}$ and $c_v = d_v = 10^{-3}$. The conditional posterior distributions of ν and and τ_v are

$$\nu \mid \mathbf{v}, \tau_v \sim N\left(\frac{N_v \tau_v \overline{v}_\cdot + b_\nu a_\nu}{N_v \tau_v + b_\nu}\ ,\ \left(N_v \tau_v + b_\nu\right)^{-1}\right) \tag{6.25a}$$

$$\tau_v \mid \mathbf{v}, \nu \sim Gamma\left(c_v + \frac{N_v}{2}, d_v + \frac{\sum_{j=1}^{N_v}\left(v_j - \nu\right)^2}{2}\right) \tag{6.25b}$$

where $\mathbf{v} = (v_j\ \forall\ j)$, N_v is the number of reanalyses and $\overline{v}_\cdot = \sum v_j / N_v$.

So now we have one framework that relates the models to the actual climate, and another that relates the reanalyses to the actual climate. This leaves us with a problem in combining evidence, since we effectively have two competing estimates of the actual historical climate

$$y_H \mid \mathbf{x} = \mu + \Delta_H \quad \text{and} \quad y_H \mid \mathbf{v} = \nu + \Delta_V$$

Provided that the two estimates are judged to be independent, then we can consider the estimate from reanalysis to be simply a normalised likelihood based only on observations (Berger, 1985, pages 271-277). In that case, a simple application of Bayes' theorem provides a solution. The posterior estimate of y_H from the climate models can be treated as the prior for the actual climate so that

$$\Pr\left(y_H \mid \mathbf{x}, \mathbf{v}\right) \propto \Pr\left(\mathbf{v} \mid y_H\right) \Pr\left(y_H \mid \mathbf{x}\right)$$

where $\Pr\left(\mathbf{v} \mid y_H\right)$ is the likelihood of the reanalyses given the actual climate, and $\Pr\left(y_H \mid \mathbf{x}\right)$ is the posterior distribution of the actual climate given the models. This is analagous to Equation D.2, or the integral over the final two terms in Equation 6.19 with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Therefore, to fit the full framework using reanalysis data, first

1. Specify the scaling factor κ_v;

2. Obtain N samples from the joint posterior of the reanalysis parameters (ν, σ_v^2), as described above;

then proceed exactly as described in the previous section, substituting samples of ν for z and $\sigma_{\Delta_V}^2 = \kappa_v^2 \sigma_v^2$ for σ_z^2.

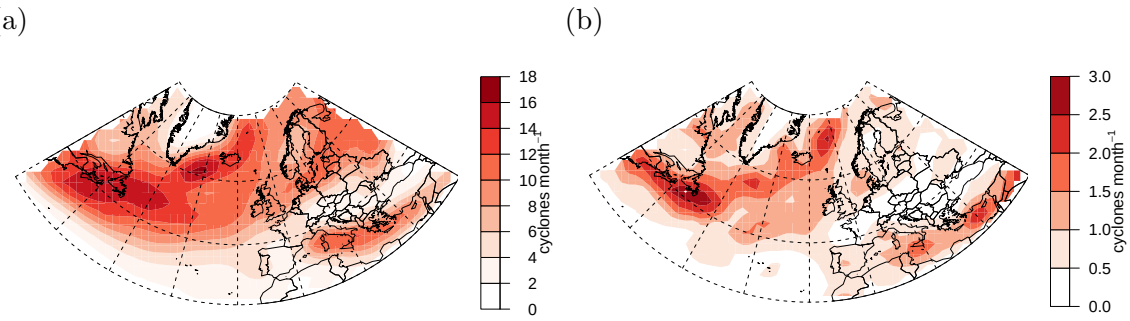

Figure 6.5.: (a) The posterior mean of the expected value of the reanalyses (ν); and (b) the square root of the posterior mean spread in the reanalyses ($\sqrt{\sigma_v^2}$)

6.8. Results

The analysis in Chapter 5 showed that there was no robust evidence of an emergent relationship in the track density of extra-tropical cyclones. While there did appear to be an emergent relationship present in the full ensemble, a closer inspection revealed that it was influenced by three outlying models. Therefore, the analysis of the North Atlantic storm track will be brief and will focus on the exchangeable ensemble which excludes the outlying models.

Bracegirdle and Stephenson (2013) studied the effect of emergent constraints in the CMIP5 multi-model ensemble on near surface temperature in the Arctic. The mechanisms underlying the emergent relationship noted in both the CMIP3 and CMIP5 ensembles are at least partially understood in terms of the simulation of sea ice thickness and extent (Holland and Bitz, 2003). There is reason to expect an emergent relationship of negative sign over most of the Arctic, particularly where sea ice forms on a seasonal rather than a permanent basis. This dataset provides an excellent opportunity to compare projection methods that include emergent relationships. In particular, the framework proposed here will be contrasted with the framework of Tebaldi et al. (2005).

6.8.1. The North Atlantic storm track

Climatologies of extra-tropical cyclones are usually derived from reanalysis data rather than directly from observations. Only version 2 of the NOAA Twentieth Century Reanalysis (Compo et al., 2011) currently includes any kind of uncertainty estimate. However, that dataset is based only on surface pressures, and so is unsuitable for storm tracking which is often performed on the relative vorticity above the boundary layer. Therefore, multiple reanalysis datasets are combined in order to estimate the observational uncertainty, as described in Section 6.7. The available global reanalysis datasets are summarised in Table 6.1. The ERA-15 and ERA-40

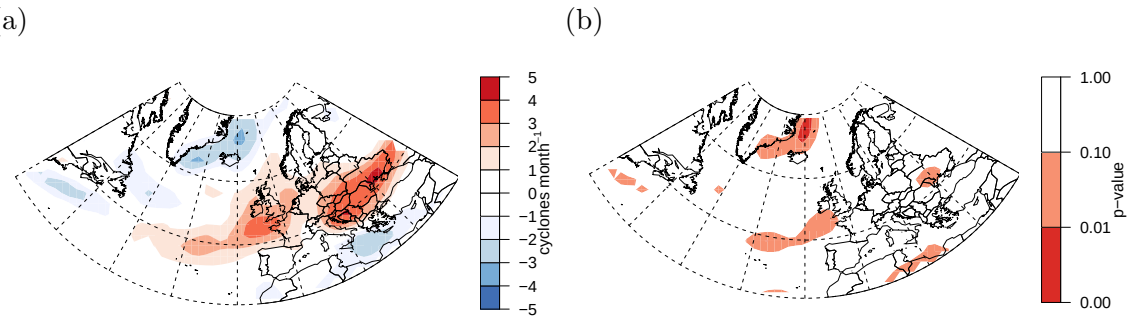

Figure 6.6.: (a) The posterior mean of the bias between the expected values of the models and the reanalyses ($\mu - \nu$); and (b) the p-value of the expected value of the reanalyses (ν) in the the posterior distribution of the historical climates of the models included in the exchangeable ensemble.

Centre	Reanalysis	Record length	Resolution	Assimilation scheme
ECMWF	ERA-15	1979/01 - 1993/12	2.5 x 2.5 L31	ECMWF Operational
ECMWF	ERA-40	1957/01 - 2002/12	1.1 x 1.1 L60	3DVAR
ECMWF	ERA-Interim	1979/01 -	0.8 x 0.8 L60	4DVAR
JMA	JRA-25	1979/01 - 2004/12	1.1 x 1.1 L40	3DVAR
JMA	JRA-55	1958/01 - 2012/12	0.5 x 0.5 L60	4DVAR
NCEP	Climate Forecast System Reanalysis	1979/01 - 2010/12	0.5 x 0.5 L64	3DVAR
NCEP-NCAR	Reanalysis (R-1)	1948/01 -	2.5 x 2.5 L28	3DVAR
NCEP-DOE	Reanalysis (R-2)	1979/01 -	2.5 x 2.5 L28	3DVAR
NASA	MERRA	1979/01 -	0.5 x 0.7 L72	GEOS IAU
NOAA	20th Century Reanalysis Version 1	1908/01 - 1958/12	2.0 x 2.0 L28	Ensemble Kalman Filter
NOAA	20th Century Reanalysis Version 2	1871/01 - 2012/12	2.0 x 2.0 L28	Ensemble Kalman Filter

Table 6.1.: Details of the available global reanalyses. Resolution is in degrees, Lxx indicates number of vertical levels. The data in this table were gathered from the references given in the text and supplemented by information from Dee et al. (2014).

products have been superseded by ERA-Interim (Dee et al., 2011). Similarly, the NCEP-NCAR and NCEP-DOE reanalysis products have been superseded by the NCEP Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010). The JRA-55 dataset was not available in time for inclusion in this thesis, therefore the JRA-25 (Onogi et al., 2007) dataset was included instead. This leaves four reanalysis products for inclusion in the analysis: ERA-Interim, JRA-25, NCEP CFSR and NASA MERRA (Rienecker et al., 2011). The cyclone climatologies from these datasets were compared by Hodges et al. (2011) who concluded that the storm tracks were well represented in all four datasets in the Northern Hemisphere. Therefore, it seems reasonable to judge that these four datasets are exchangeable and can be used as an alternative to observations, as described in Section 6.7. The posterior mean estimates of the expectation (ν) and spread ($\sqrt{\sigma_v^2}$) of the reanalyses are shown in

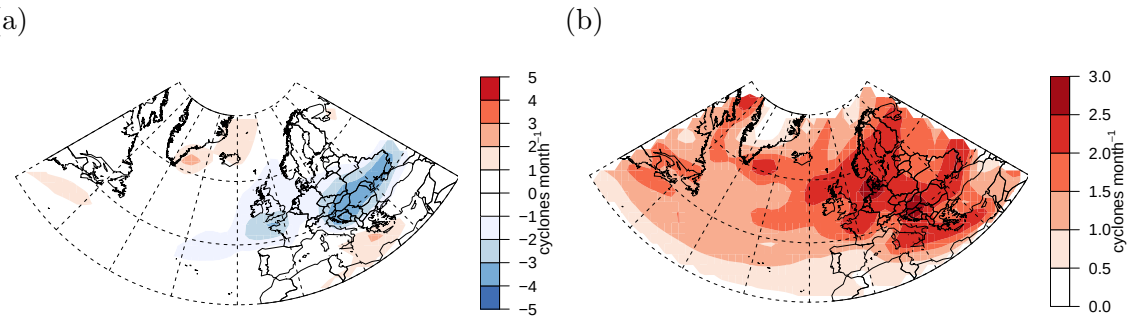

Figure 6.7.: (a) The posterior mean of the historical discrepancy (Δ_H); and (b) the square root of prior uncertainty about the historical discrepancy ($\sqrt{\sigma^2_{\Delta_H}}$).

Figure 6.5. The spread is largest in the most active part of the storm track between Newfoundland and Iceland and beyond. This is unsurprising since Hodges et al. (2011) noted that although agreement was very good for the strongest storms, there was more variation in the weaker systems which dominate the cyclone counts.

The natural variability in the Earth system, quantified by σ^2_{Ha} and σ^2_{Fa}, must also be estimated. Ideally the historical uncertainty σ^2_{Ha} would be estimated from a long time series of observations. However, annual track density statistics were not available, so the estimates of the internal variability simulated by the models σ^2_H and σ^2_F were substituted instead.

Since no robust evidence of an emergent constraint was found in Chapter 5, λ is set to 0 unless otherwise stated. For this analysis, the scaling factor κ is fixed at 1, implying that the models are judged to be exchangeable with the actual climate. This judgement is perhaps too strong, but it is likely to be the most common choice in practice. At the very least, $\kappa = 1$ satisfies the intuition that our uncertainty about the actual climate response is at least as great as the spread of responses simulated by the models. However, if the observed climate lies well outside of the range of historical climates simulated by the models, then it is difficult to conclude that the models are exchangeable with the actual climate. The observed climate, as approximated by the expected value of the reanalyses (ν), lies within the spread of historical climates simulated by the models over most of the North Atlantic domain (Figure 6.6b), despite large biases between the historical climate of the ensemble, and the reanalyses. (Figure 6.6a). Therefore, the judgement that $\kappa = 1$ seems reasonable.

The posterior mean of the historical discrepancy $\Delta_H = y_H - \mu$ is small over most of the study area (Figure 6.7a). Note the change in sign between the discrepancy Δ_H, and the bias $\mu - \nu$ (Figure 6.6a). The difference in magnitude between the discrepancy and the bias ($y_H - \nu$) is equal to the shrinkage of the estimate of the actual climate away from the observations and towards the ensemble (Figure 6.8a).

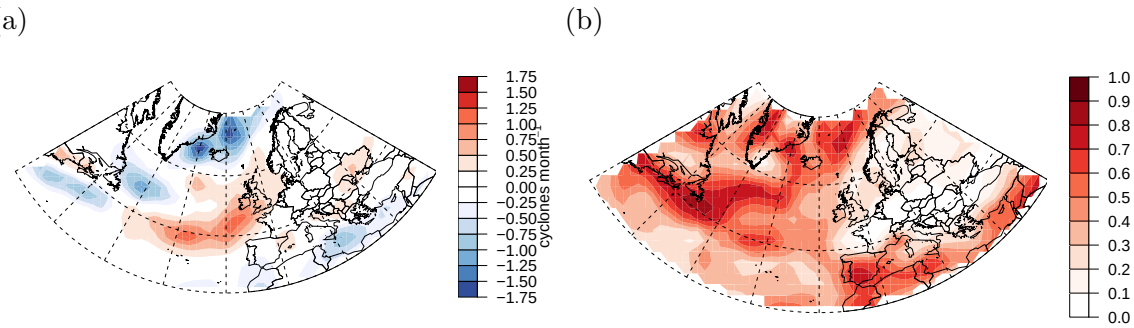

Figure 6.8.: The posterior means of (a) the shrinkage of actual climate y_H away from the expected climate of the reanalyses and towards the expected climate of the ensemble ($y_H - \nu$); (b) the information ratio I (Section 6.4). $I > 0.5$ indicates that y_H is estimated to lie closer to the expected climate of the models μ than the mean of the reanalyses ν.

Shrinkage of more than one cyclone per month occurs near Newfoundland, Iceland, and south west of the United Kingdom. Large shrinkages occur wherever large biases $\mu - \nu$ (Figure 6.6a) coincide with large values of the information ratio I (Figure 6.8b). So despite large biases, very little shrinkage occurs over Eastern Europe, because the models are not at all informative compared to the reanalyses ($I < 0.1$). Whereas near Newfoundland and Iceland, the spread in the reanalyses is large (Figure 6.5b) compared to the prior uncertainty about the historical discrepancy (Figure 6.7b), so large shrinkages occur. This is likely to be due differences in the number of weak cyclones identified in the reanalyses (Hodges et al., 2011).

Since no emergent constraint is included, the projected response of the actual climate $y_F - y_H$ is simply the expected response of the ensemble (Figure 6.9). The standard error of the projected response is of a similar magnitude to the response over most of the study region (Figure 6.10a). Only in the Mediterranean basin is the signal strong enough to be considered significant at the 10% level (not shown) once the uncertainty about the discrepancy between the ensemble and the actual climate is accounted for. The standard error of the climate response that we might experience is more than 40%, and often more than 60%, larger than the standard error of the actual climate response over most of the study region (Figure 6.10b). This agrees with the conclusion of Chapter 3 where we found that the internal variability was large compared to the differences between the models over much of the North Atlantic.

In Chapter 5, no evidence of a robust emergent relationship was found. The effect of including an emergent relationship in the projected response is small over most of the region (Figure 6.11a). However, differences of more than one cyclone per month occur where large incidental emergent constraint estimates (Figure 5.5a) coincide with large historical discrepancies (Figure 6.7a). In the full ensemble, a weak emergent relationship was noted in the sub-tropics due to the influence of three outlying mod-

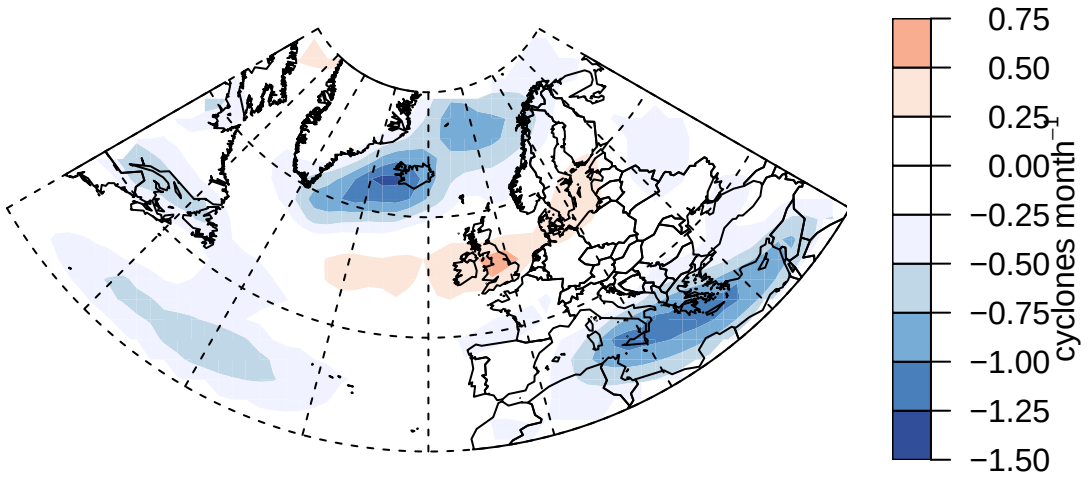

Figure 6.9.: The posterior mean of the actual climate response y_R.

(a) (b)

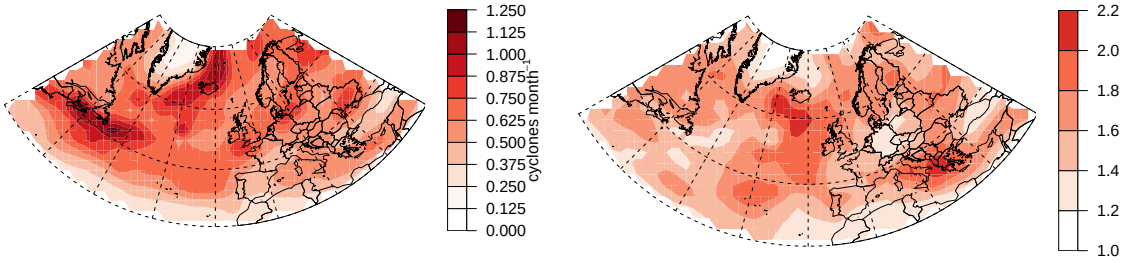

Figure 6.10.: (a) The standard error of the actual climate response y_R; and (b) the ratio of the standard error of the climate response that we might experience due to natural variability y_{Ra} to that of the actual climate response y_R.

els. However, the projected response from the full ensemble, including an emergent constraint, is not as different from the actual response projected by the exchangeable ensemble with no emergent constraint as might be expected (Figure 6.11b). This is due to a combination of factors. First of all, the expected responses of the models differs between the two ensembles (Figure 6.12a). The three influential models noted in Chapter 5 all simulate strong decreases in cyclone activity in the subtropics, causing a negative bias in the expected response of the ensemble β. This partially cancels the effect of including the emergent relationship (Figure 6.12b). As a result, the effect of predicting from the full ensemble including the spurious emergent constraint is not large. However, cancellation in this way might not always occur. Therefore, care must still be taken to understand the physical mechanisms behind any apparent emergent relationships. Only when a relationship is well understood should it be used for projection.

(a) (b)

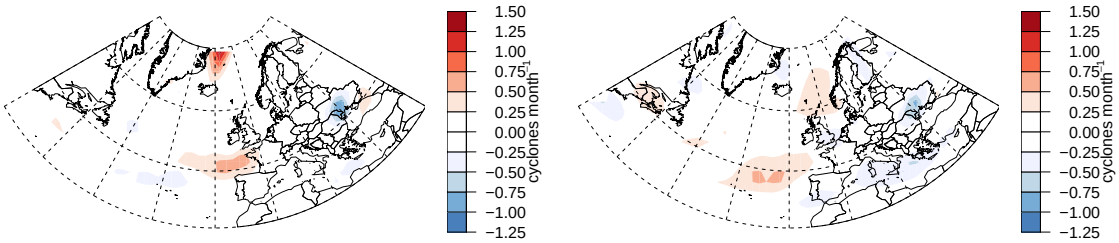

Figure 6.11.: The difference between the posterior mean of the projected response y_R estimated (a) with an emergent constraint from the exchangeable ensemble, and (b) with an emergent constraint from the full ensemble, and the estimate without an emergent constraint from the exchangeable ensemble in Figure 6.9.

(a) (b)

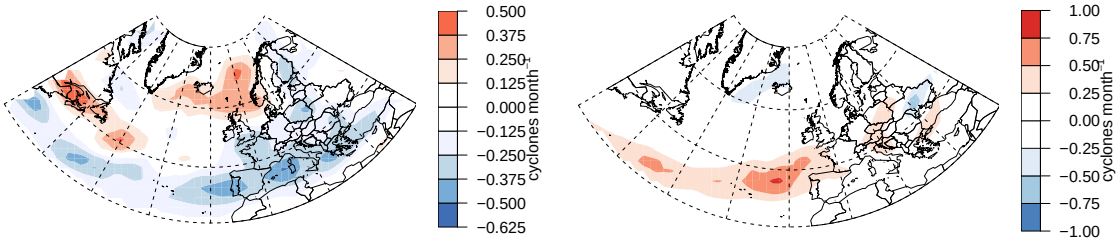

Figure 6.12.: The posterior mean of (a) the difference between the expected response β of the full ensemble and that of the exchangeable ensemble; and (b) the difference between the projected response y_R including an emergent constraint, and the expected response of the ensemble β, both estimated from the full ensemble.

6.8.2. Arctic near surface temperature

An extended version of the CMIP5 surface temperature dataset analysed by Bracegirdle and Stephenson (2013) is considered in this section. The mean climates over 30 winters (December-January-February) are compared between December 1975 and January 2005 from the historical scenario, and between December 2069 and January 2099 from the RCP4.5 scenario. The five year shift in the historical period compared to Bracegirdle and Stephenson (2013) provides slightly better compatibility with the latest observation and reanalysis datasets. Several of these datasets begin in 1979 when satellite observations become prevalent. A total of 216 runs from 37 CMIP5 models are included in the full ensemble, 128 runs of the historical scenario and 88 of the RCP4.5 scenario. The number of runs available from each model is listed in Table 6.2.

The full ensemble was thinned in order to satisfy the judgement of exchangeability between the model outputs. Unless otherwise stated, the same model is included from each modelling centre as in Chapter 4. Several models were excluded from the analysis of the North Atlantic storm track due to badly displaced storm tracks. Those models are reinstated for the analysis of Arctic surface temperature, although

Table 6.2.: Number of realisations available from each model for the historical and future scenarios. Models highlighted in red are included in the exchangeable ensemble.

Modelling centre	Model	Runs	
		Historical	RCP4.5
		R_{Hm}	R_{Fm}
CSIRO-BOM	ACCESS1.0	1	1
CSIRO-BOM	ACCESS1.3	3	1
BCC	BCC-CSM1.1	3	1
BCC	BCC-CSM1.1(m)	3	1
BNU	BNU-ESM	1	1
CCCMA	CanESM2	5	5
NCAR	CCSM4	6	6
NSF-DOE-NCAR	CESM1(BGC)	1	1
NSF-DOE-NCAR	CESM1(CAM5)	3	3
NSF-DOE-NCAR	CESM1(WACCM)	4	1
CMCC	CMCC-CM	1	1
CMCC	CMCC-CMS	1	1
CNRM-CERFACS	CNRM-CM5	10	1
CSIRO-QCCCE	CSIRO-Mk3.6.0	10	10
ICHEC	EC-EARTH	8	9
LASG-CESS	FGOALS-g2	5	1
FIO	FIO-ESM	3	3
NOAA GFDL	GFDL-CM3	5	1
NOAA GFDL	GFDL-ESM2G	1	1
NOAA GFDL	GFDL-ESM2M	1	1
NASA GISS	GISS-E2-H	6	5
NASA GISS	GISS-E2-R	6	6
NIMR/KMA	HadGEM2-AO	1	1
MOHC	HadGEM2-CC	3	1
MOHC	HadGEM2-ES	4	4
INM	INM-CM4	1	1
IPSL	IPSL-CM5A-LR	6	4
IPSL	IPSL-CM5A-MR	3	1
IPSL	IPSL-CM5B-LR	1	1
MIROC	MIROC5	5	3
MIROC	MIROC-ESM	3	1
MIROC	MIROC-ESM-CHEM	1	1
MPI-M	MPI-ESM-LR	3	3
MPI-M	MPI-ESM-MR	3	3
MRI	MRI-CGCM3	3	1
NCC	NorESM1-M	3	1
NCC	NorESM1-ME	1	1
Total		128 (63)	88 (40)

several are subsequently excluded in the thinning process. Results from some additional models are included that were not contained in the cyclone track density data analysed by Zappa et al. (2013b) and in the preceding chapters. The complete list of models and details of their major components are given in Table 6.3. Three models were submitted from the combined efforts of the NSF-DOE-NCAR.

Modelling centre	Model	Atmosphere	Atmosphere res.	Ocean	Ocean res.	Sea ice	Land surface
CSIRO-BOM	ACCESS1.0	HadGEM2 (r1.1)	1.9 x 1.9 L38	MOM4.1	1.0 x 1.0 L50	CICE4.1	MOSES2.2
CSIRO-BOM	ACCESS1.3	UKMO GA 1.0	1.9 x 1.9 L38	MOM4.1	1.0 x 1.0 L50	CICE4.1	CABLE
BCC	BCC-CSM1.1	BCC_AGCM2.1	2.8 x 2.8 L26	MOM4-L40	1.0 x 1.0 L40	GFDL SIS	BCC-AVIM1.0
BCC	BCC-CSM1.1(m)	BCC_AGCM2.1	1.1 x 1.1 L26	MOM4-L40	1.0 x 1.0 L40	GFDL SIS	BCC-AVIM1.0
GCESS	BNU-ESM	CAM3.5	2.8 x 2.8 L26	MOM4.1	1.0 x 1.0 L50	CICE4.1	CoLM+BNUDGVM
CCCMA	CanESM2	CanAM4	2.8 x 2.8 L35	CanOM4	1.4 x 1.4 L40	Included	CLASS 2.7; CTEM
NCAR	CCSM4	CAM4	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
NSF-DOE-NCAR	CESM1(BGC)	CAM4	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
NSF-DOE-NCAR	CESM1(CAM5)	CAM5	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
NSF-DOE-NCAR	CESM1(WACCM)	WACCM4	2.5 x 2.5 L66	POP2	1.0 x 1.0 L60	CICE4	CLM4
CMCC	CMCC-CM	ECHAM5	0.8 x 0.8 L31	OPA8.2	2.0 x 2.0 L31	LIM2	N / A
CMCC	CMCC-CMS	ECHAM5	1.9 x 1.9 L95	OPA8.2	2.0 x 2.0 L31	LIM2	N / A
CNRM-CERFACS	CNRM-CM5	ARPEGE-Climat	1.4 x 1.4 L31	NEMO	0.7 x 0.7 L42	Gelato5	SURFEX
CSIRO-QCCCE	CSIRO-Mk3.6.0	Included	1.9 x 1.9 L18	MOM2.2	1.9 x 1.9 L31	Included	Included
EC-EARTH	EC-EARTH	IFS c3 1r1	1.1 x 1.1 L62	NEMO_ecmwf	1.0 x 1.0 L31	LIM2	HTESSEL
LASG-CESS	FGOALS-g2	GAMIL2	2.8 x 2.8 L26	LICOM2	1.0 x 1.0 L30	CICE4-LASG	CLM3
FIO	FIO-ESM	CAM3.0	2.8 x 2.8 L26	POP2	1.0 x 1.0 L40	CICE4	CLM3.5
NOAA GFDL	GFDL-CM3	Included	2.5 x 2.5 L48	MOM4.1	1.0 x 1.0 L50	SIS	Included
NOAA GFDL	GFDL-ESM2G	AM2.1	2.5 x 2.5 L24	GOLD	1.0 x 1.0 L63	SIS	Included
NOAA GFDL	GFDL-ESM2M	AM2.1	2.5 x 2.5 L60	MOM4.1	1.0 x 1.0 L50	SIS	Included
NASA GISS	GISS-E2-H	Inlcuded	2.5 x 2.5 L40	HYCOM	1.0 x 1.0 L26	Included	Included
NASA GISS	GISS-E2-R	Inlcuded	2.5 x 2.5 L40	Russell	1.0 x 1.0 L32	Included	Included
NIMR/KMA	HadGEM2-AO	HadGAM2	1.9 x 1.9 L60	Included	1.9 x 1.9	Included	Included
MOHC	HadGEM2-CC	HadGAM2	1.9 x 1.9 L60	Included	1.9 x 1.9	Included	Included
MOHC	HadGEM2-ES	HadGAM2	1.9 x 1.9 L38	Included	1.0 x 1.0 L40	Included	Included
INM	INM-CM4	Included	2.0 x 2.0 L21	Included	1.0 x 1.0 L40	Included	Included
IPSL	IPSL-CM5A-LR	LMDZ5	3.8 x 3.8 L39	NEMO	2.0 x 2.0 L31	Included	Included
IPSL	IPSL-CM5A-MR	LMDZ5	2.5 x 2.5 L39	NEMO	2.0 x 2.0 L31	Included	Included
IPSL	IPSL-CM5B-LR	LMDZ5	3.8 x 3.8 L39	NEMO	2.0 x 2.0 L31	Included	Included
MIROC	MIROC5	MIROC-AGCM6	1.4 x 1.4 L40	COCO4.5	1.4 x 1.4 L50	Included	MATSIRO
MIROC	MIROC-ESM	MIROC-AGCM	2.8 x 2.8 L80	COCO3.4	1.4 x 1.4 L44	Included	MATSIRO
MIROC	MIROC-ESM-CHEM	MIROC-AGCM	2.8 x 2.8 L80	COCO3.4	1.4 x 1.4 L44	Included	MATSIRO
MPI-M	MPI-ESM-LR	ECHAM6	1.9 x 1.9 L47	MPIOM	1.5 x 1.5 L40	Included	JSBACH
MPI-M	MPI-ESM-MR	ECHAM6	1.9 x 1.9 L95	MPIOM	0.4 x 0.4 L40	Included	JSBACH
MRI	MRI-CGCM3	MRI-AGCM3.3	1.1 x 1.1 L48	MRI.COM3	1.0 x 1.0 L50	MRI.COM3	HAL
NCC	NorESM1-M	CAM4-Oslo	2.5 x 2.5 L26	NorESM-Ocean	1.1 x 1.1 L53	CICE4	CLM4
NCC	NorESM1-ME	CAM4-Oslo	2.5 x 2.5 L26	NorESM-Ocean	1.1 x 1.1 L53	CICE4	CLM4

Table 6.3.: Structural details of the 37 CMIP5 models included in the analysis of Arctic surface temperature. Models highlighted in red are included in the exchangeable ensemble. Atmosphere and ocean resolution are in degrees and Lxx indicates the number of vertical levels. Details included in this table were gathered from the metadata included in the model outputs and supplemented using information from Table 9.A.1 of Flato et al. (2013).

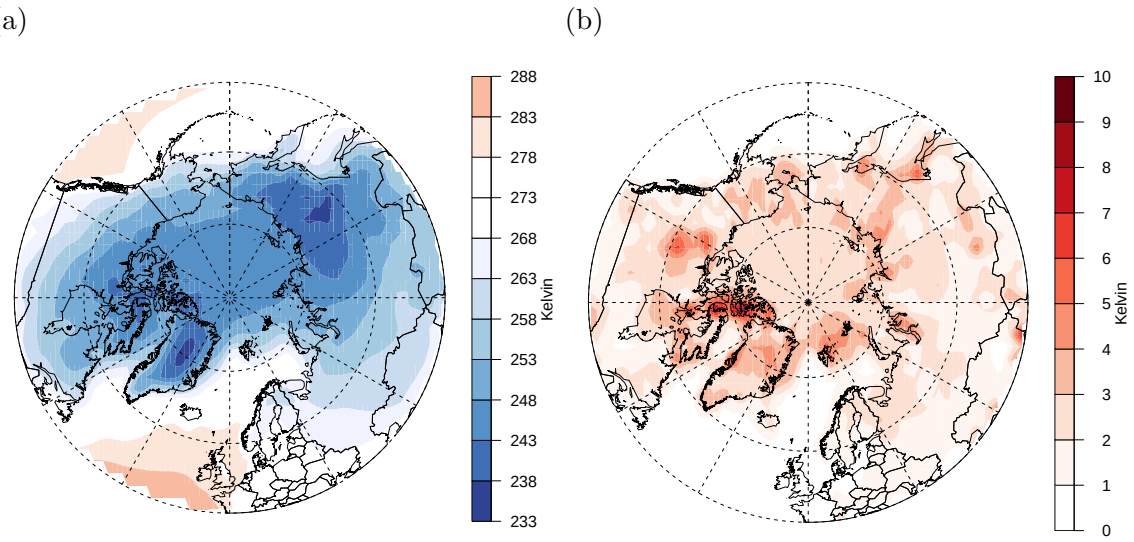

Figure 6.13.: (a) The posterior mean of the expected value of the reanalyses (ν); and (b) the square root of the posterior spread in the reanalyses ($\sqrt{\sigma_v^2}$)

The CESM1(CAM5) variant was selected as it includes a more recent version of the CAM atmosphere model. Two variants of the NCC model were available for surface temperature. The basic NorESM1-M version was selected since many models do not yet include the interactive ocean bio-geochemistry module that differentiates the more complex variant. The ACCESS models supersede the CSIRO-Mk3.6.0 model, however all of the major components in the ACCESS models are borrowed from other models. Therefore, none of the models submitted by CSIRO were included. The two models submitted from the CMCC are both based on an old atmosphere component and a very old ocean component. They also lack a full land surface model, therefore neither model was included. The BNU-ESM and FIO-ESM models were also excluded since they use outdated and low resolution versions of the CAM atmosphere included in the CESM1 model. The NCAR CCSM4 model has also been superseded by the CESM1 model, and so was not included either. This leaves an ensemble of 104 runs from 15 models, 63 runs from the historical scenario and 41 from the RCP4.5 scenario. The models and runs included in the thinned ensemble are indicated in Table 6.2. All the analysis that follows relates to the thinned ensemble unless otherwise stated.

Estimating the observation and sampling uncertainty

Unlike extra-tropical cyclones, surface temperature is often studied directly from observational data rather than from reanalyses. Several analyses of global temperature observations exist, however none are entirely suitable for use here. The NASA GISS dataset (Hansen et al., 2010) does not include any explicit estimate of the uncertainty in the analysis. The NOAA MLOST dataset (Vose et al., 2012)

does include a basic estimate of the analysis uncertainty. However, the uncertainty estimate is limited to the effect of sparsity in the observations, and SST bias adjustments. The HadCRUT4 analysis (Morice et al., 2012) includes a sophisticated assessment of the uncertainty underlying the observations. However, the analysis is provided as anomalies from a reference period rather than as absolute temperatures. Furthermore, no spatial infilling takes place, so coverage is poor in the Arctic where observations are very sparse. Since none of the observational analyses have both the spatial coverage and uncertainty assessment required for combination with model data, reanalysis data is once again used. The same four reanalyses are used as in Section 6.8.1, namely: ERA-Interim, JRA-25, NASA MERRA and NCEP CFSR. The posterior mean estimates of the expectation (ν) and spread ($\sqrt{\sigma_v^2}$) of the reanalyses are shown in Figure 6.13. The reanalyses agree well over the oceans in the mid-latitudes, because the large heat capacity of the ocean means that surface temperature is quite stable there. In the Arctic (above 66N), the spread between the reanalyses is greater since observations are very sparse, and the representation of sea ice and other complex physical processes are critical.

For consistency with the cyclone track density analysis, the same approach is used to estimate the sampling uncertainty σ_{Ha}^2 and the natural variability σ_{Fa}^2. The estimates of the internal variability simulated by the models, σ_H^2 and σ_F^2, are substituted for the sampling uncertainty and natural variability.

Fitting the full framework

The default vague priors were used throughout and no difficulties were encountered. Once again, the Gibbs samplers converged very quickly with little or no burn-in period evident. However, small but significant autocorrelation was visible for up to 2,000 samples, particularly in the expectation climate parameters μ and ν. Therefore, only every 2,000th sample was retained, after a burn-in period of 20,000 samples. Posterior estimates were based on 10,000 samples after thinning. In this configuration, each grid box required approximately 143 seconds of compute time using the Fortran implementation of the framework. Projecting surface temperature in the Northern Hemisphere above 45N involved fitting the model to 2,592 grid boxes. This required a total of just over 100 hours of compute time, which was achieved in a little over one day by splitting the work between multiple CPU cores. Once again, we adopt the default assumption that the prior uncertainty about the ensemble discrepancy is equal to the model uncertainty ($\kappa = 1$).

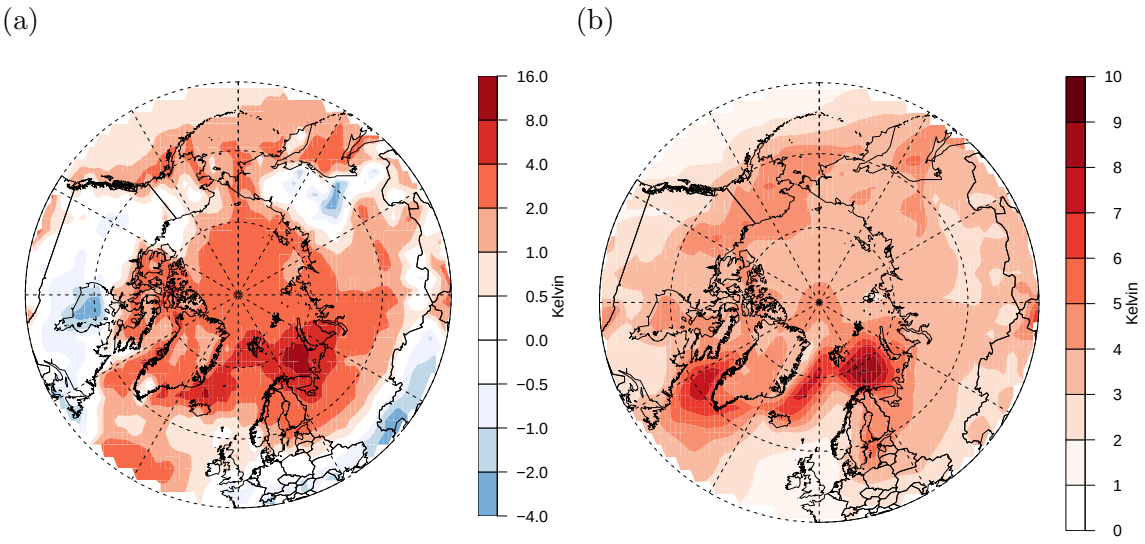

Figure 6.14.: (a) The posterior mean of the historical discrepancy (Δ_H); and (b) the square root of the prior uncertainty about the historical discrepancy ($\sqrt{\sigma^2_{\Delta_H}}$).

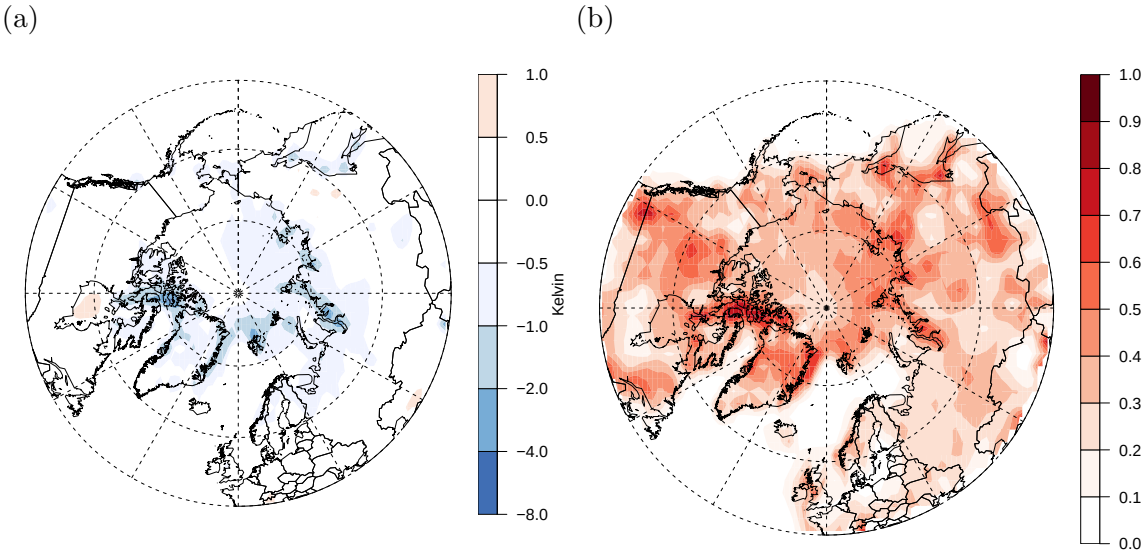

Figure 6.15.: The posterior means of (a) the shrinkage of the observed climate towards the expected climate of the ensemble ($y_H - \nu$); (b) the information ratio I (Section 6.4). $I > 0.5$ indicates that y_H is estimated to lie closer to the expected climate of the models μ than the mean of the reanalyses ν.

Combining models and observations

The historical discrepancy Δ_H between the expected climate of the ensemble μ and the estimate of the actual climate y_H is fairly uniform at 2-4K or less over most of the Arctic Ocean (Figure 6.14a). However, differences exceeding 6K occur in the Greenland Sea and Denmark Strait, and in the Barents Sea they may exceed 10K. As in previous studies (e.g., Bracegirdle and Stephenson, 2012, Figure 5d), the models are generally too cold compared to the actual climate. The expected climate of the reanalyses ν lies within the spread of the historical model climates at all but

a handful of scattered grid points (not shown). Therefore, the judgement that the prior uncertainty about the historical discrepancy $\sigma^2_{\Delta_H}$ is equal to the model spread σ^2_α ($\kappa = 1$) seems reasonable. The prior uncertainty about the historical discrepancy $\sigma^2_{\Delta_H}$ is fairly uniform above 60N. The exception is the North Atlantic, where the prior uncertainty is much larger along the ice edge between the Labrador sea and the Barents sea (Figure 6.14b).

The information ratio is small ($I < 0.1$) over the mid-latitude oceans where the reanalyses are in good agreement (Figure 6.15b). Over the land and in the Arctic it is larger ($0.2 < I < 0.4$), but rarely exceeds 0.5. Therefore, the shrinkage of the estimate of the actual climate y_H away from the reanalyses ν and towards the models μ is generally small, less than 1K over most of the Arctic and less than 0.5K elsewhere (Figure 6.15a). Given the known deficiencies in the models' representation of the cryosphere, we might not believe large adjustments towards the ensemble climate. However, observation uncertainty in the Arctic is large, so some adjustment seems appropriate. If we held strong beliefs that the models should be even less informative for the historical climate, then these could be incorporated by setting the scaling factor $\kappa > 1$.

The projected temperature response

The projected temperature response $y_R = y_F - y_H$ is fairly uniform at high latitudes, 4-6K between 60-75N, and 6-8K over most of the Arctic Ocean (Figure 6.16). The strongest response occurs in the north of the Barents Sea, east of Svalbard, where the response may exceed 12K. The expected response of the models is reinforced by a weak positive emergent relationship in this region (Figure 6.18a). The standard error of the projected response tends to increase with latitude (Figure 6.17a). At high latitudes the models are less informative for the climate response, due to the difficulty in representing the complex processes associated with snow and ice cover. Like the expected response, the standard error is also largest in the north of the Barents Sea. The standard error of the climate response that we might experience y_{Ra} is less than 5% greater than that of the expected climate response over most of the Arctic and Atlantic Oceans (Figure 6.17b). Over land where temperatures are more variable, the posterior standard error for the response of the actual climate increases by 5-40% when natural variability is included.

Over most of the Arctic Ocean, the projected response incorporating an emergent relationship is at least 0.5-1.0K lower than a comparable projection without an emergent constraint (Figure 6.19a). In some regions, the projected response including the emergent constraint may be more than 2K lower. The largest differences tend to coincide with the strongest emergent constraints (Figure 6.18a), since the historical

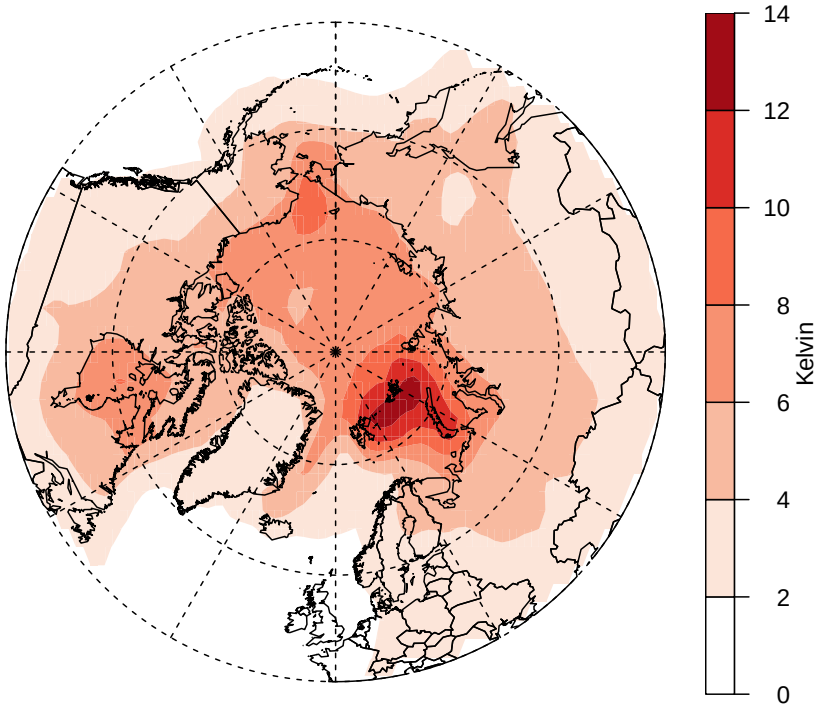

Figure 6.16.: The posterior mean of the actual climate response y_R.

(a) (b)

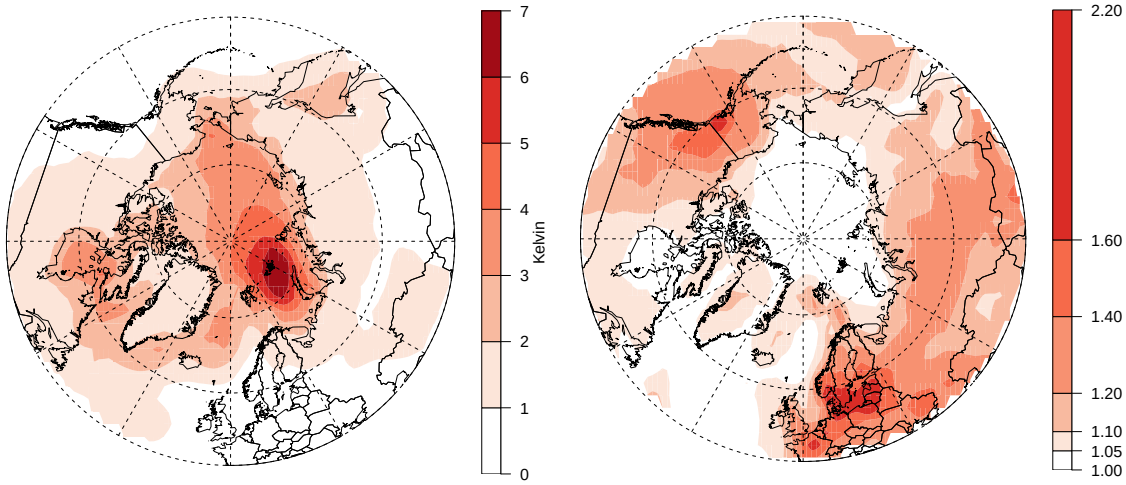

Figure 6.17.: (a) The standard error of the actual climate response y_R; and (b) the ratio of the standard error of the climate response that we might experience due to natural variability y_{Ra} to that of the actual climate response y_R.

discrepancy is fairly uniform (Figure 6.14a). The projection including an emergent constraint actually predicts a more intense warming where a weak positive emergent relationship is estimated in the Barents and Kara seas.

Qualitatively, the emergent relationships estimated from the full and exchangeable ensembles are in good agreement (Figure 6.18). Quantitatively, the exchangeable ensemble estimates a much stronger negative correlation over most of the Arctic ocean,

(a) (b)

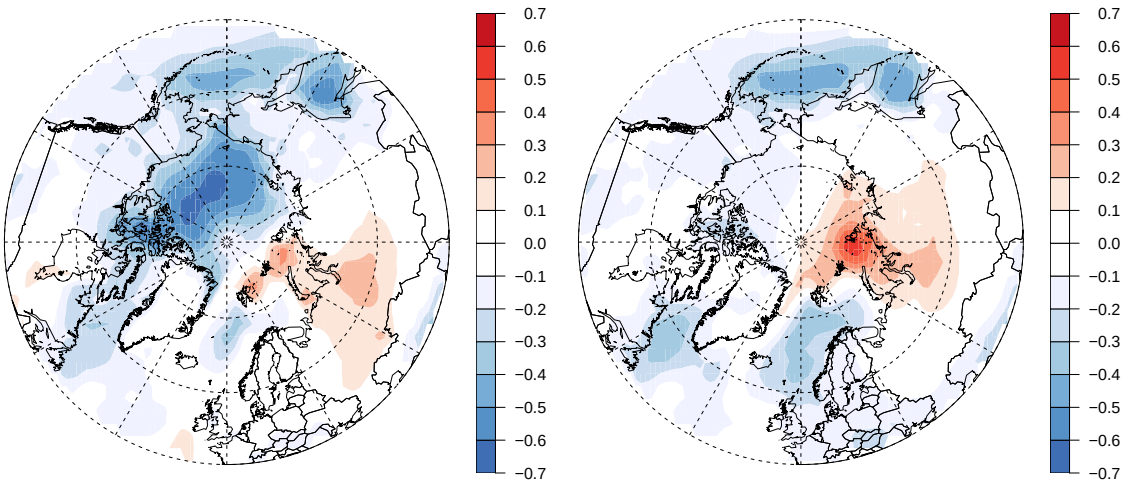

Figure 6.18.: The posterior mean of the emergent constraint λ from (a) the exchangeable ensemble; and (b) the full ensemble.

(a) (b)

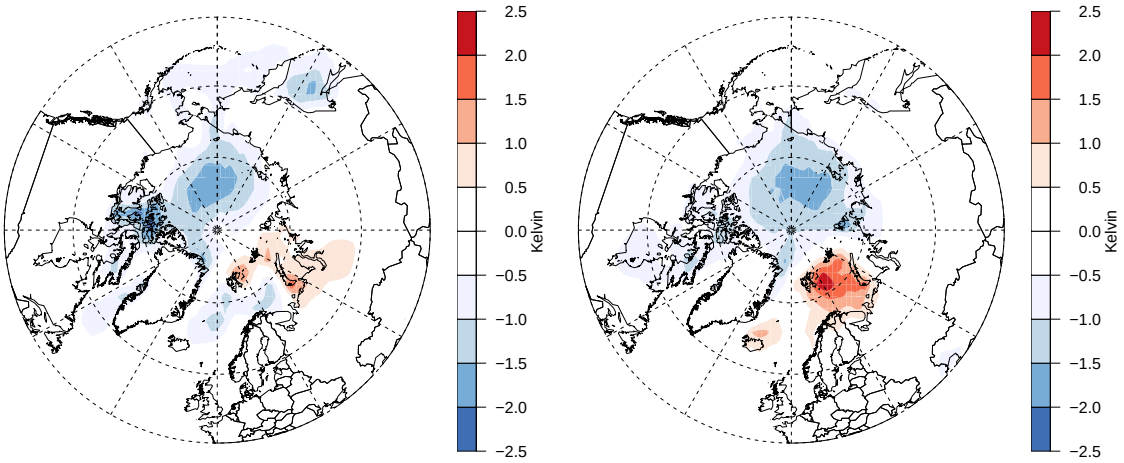

Figure 6.19.: The difference between the posterior mean estimates of the projected response y_R from (a) the exchangeable ensemble with and without an emergent constraint; and (b) the exchangeable ensemble and the full ensemble both including an emergent constraint.

and a weaker positive correlation over the Barents and Kara seas. This suggests that the choice of models included in the exchangeable ensemble may be influencing the estimation of the emergent constraint, and hence the projected response.

The difference between the projected responses of the exchangeable and full ensembles is shown in Figure 6.19b. In the Arctic ocean, the projected response from the exchangeable ensemble is up to 1.75K lower than that from the full ensemble. This agrees with the stronger negative correlation estimated by the exchangeable ensemble (Figure 6.18). However, the projected warming from the exchangeable ensemble is up to 2.5K greater than from the full ensemble in the Barents sea. The differences in the projected response can be decomposed into the difference in the

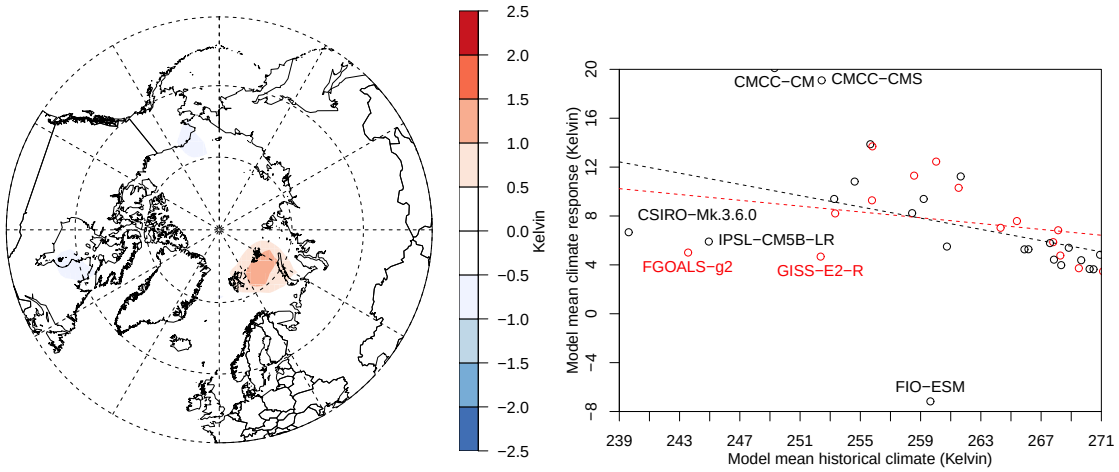

Figure 6.20.: The difference between the posterior mean estimates of the expected response of the ensemble β from the exchangeable ensemble and the full ensemble.

Figure 6.21.: As for Figure 6.22, but for a grid box in the Barents Sea (41.3E,71.3N).

expected responses of the ensembles (β) and the difference in the estimated response discrepancies (Δ_R)

$$\begin{aligned} y_r - y_r^* &= (\beta + \Delta_R) - (\beta^* + \Delta_R^*) \\ &= (\beta - \beta^*) + (\Delta_R - \Delta_R^*) \end{aligned}$$

where $*$ indicates an estimate from the full ensemble. To the north of the Barents Sea, part of the difference in the projected response y_R is due to a difference in the expected response β (Figure 6.20). Several of the models in the exchangeable ensemble simulate very strong responses in this region. Therefore the expected warming β is up to 1.5K greater than in the full ensemble. Elsewhere however, differences in the response discrepancy Δ_R dominate the change in the projected response due to the difference in the estimated emergent relationships (Figure 6.18).

In the south and west of the Barents sea, the emergent relationship in the exchangeable ensemble is influenced by the extreme cold bias and small warming simulated by FGOALS-g2 (Figure 6.21) A detailed examination of the model climates revealed that a small number of models were influencing the estimate of the emergent relationship throughout the Arctic Ocean and the Kara Sea (Figure 6.22). The CSIRO-Mk3.6.0, FGOALS-g2, and IPSL-CM5B-LR models simulate strong cold biases in the historical period, and weaker than average warming in the RCP4.5 future scenario. In contrast, MIROC-ESM and MIROC-ESM-CHEM tend to be relatively warm in the historical scenario, and simulate above average warming in the future. In the Kara sea, the two outlying groups of models induce a strong positive emergent relationship in the full ensemble, since the other models exhibit little or no correla-

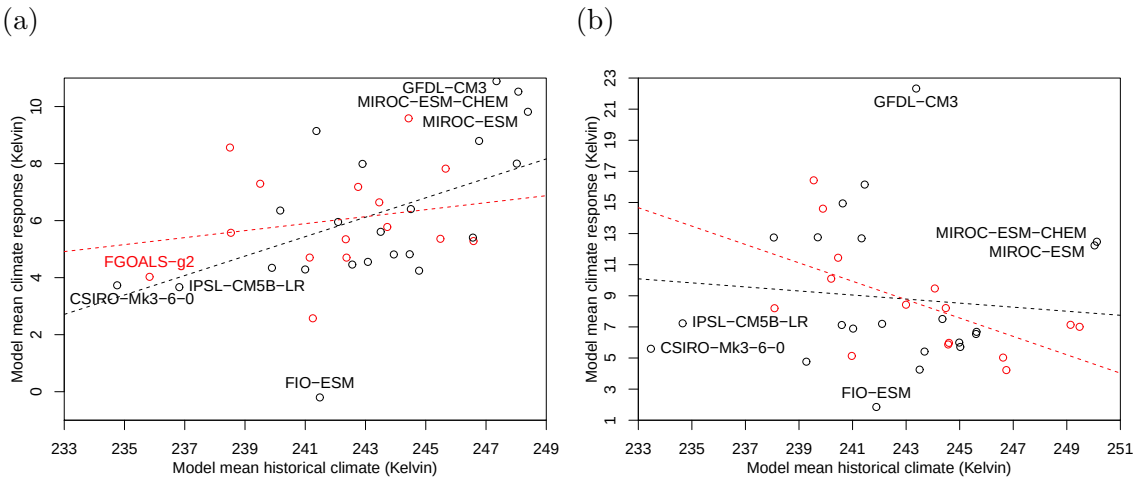

Figure 6.22.: The model mean climate response ($\overline{x}_{Fm.} - \overline{x}_{Hm.}$) plotted against the model mean historical climates ($\overline{x}_{Hm.}$) for grid boxes (a) in the Kara Sea (86.3E,73.8N) (b) in the Arctic Ocean (176.3W,76.3N). Data points in red indicate the models belonging to the exchangeable ensemble. The dashed lines are the emergent relationships estimated by ensemble regression. The black lines are computed using the full ensemble, the red lines using the exchangeable ensemble.

tion (Figure 6.22a). Over the rest of the Arctic ocean, the remaining models tend to be negatively correlated (Figure 6.22b). The outlying models act to neutralise this correlation in the full ensemble. FIO-ESM also stands out as simulating consistently weak warming across the Arctic. However, its influence is limited because its historical climate tends to agree well with the rest of the ensemble.

With the exception of FGOALS-g2, all of the models identified above as influential were excluded from the exchangeable ensemble. However, most of them were included in the analysis of Bracegirdle and Stephenson (2013). There they would have been even more influential due to the smaller ensemble size (22 models compared to 37 here). This explains the differences between the estimated emergent constraints in the full and exchangeable ensembles, and why the full ensemble more closely resembles the estimate of Bracegirdle and Stephenson (2013). In the older CMIP3 ensemble, a weak positive emergent relationship was evident over the whole of the Arctic ocean, as opposed to the negative relationship evident in CMIP5 (Bracegirdle and Stephenson, 2013). The influential models in the CMIP5 ensemble tend to induce a positive emergent relationship (Figure 6.22), and were all excluded from the exchangeable ensemble due to either low resolution or outdated model components. This might explain the reversal in sign of the emergent constraint over the Arctic Ocean between CMIP3 and CMIP5. There is reason to expect an emergent relationship anywhere that sea ice forms on a seasonal rather than permanent basis. Therefore, it seems reasonable to accept the exchangeable ensemble as more consistent with our physical understanding.

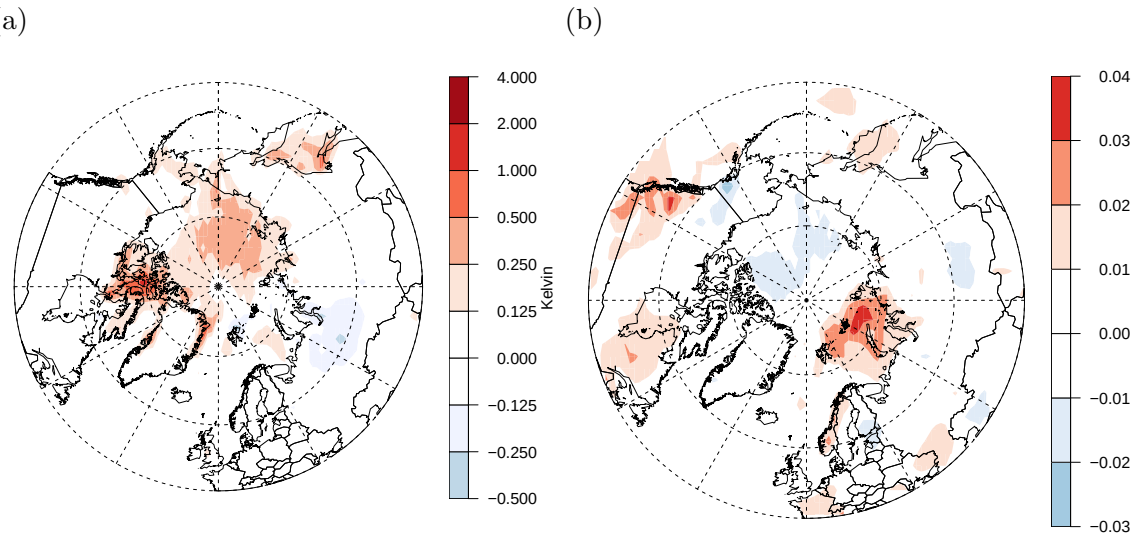

Figure 6.23.: The difference between the posterior mean estimates from the framework developed here and the maximum likelihood estimates by ensemble regression, of (a) the projected response of the actual climate (y_R); and (b) the emergent constraint (λ)

6.8.3. Comparison with other methods including emergent constraints

In this section, we briefly compare projections of near surface temperature from the framework developed in this thesis, with other frameworks that include the estimation of emergent constraints.

Ensemble regression

The ensemble regression method proposed by (Bracegirdle and Stephenson, 2012) was also fitted to the surface temperature data from the exchangeable ensemble, and the mean of the reanalyses ν used to project the response of the actual climate. The estimate of the emergent constraint λ differs by less than 0.01 over most of the study area (Figure 6.23b). This is in sharp contrast to the cyclone track density data in Chapter 5. However, the internal variability simulated by the models is small compared to the model uncertainty for surface temperature (not shown), so this agrees with the theoretical arguments in Chapter 5. The estimates of the expected value of the actual climate response y_R are also very close over most of the region (Figure 6.23a). Differences of up to 0.5K are visible where the shrinkage away from the reanalysis was large in Figure 6.15a. If the models were more informative for the actual climate, then the shrinkage would be greater and the differences between the projections would grow. Since we made the default assumption that $\kappa = 1$ in the earlier analysis, the standard errors were also similar, within 10% over more than 85% of the study area (not shown).

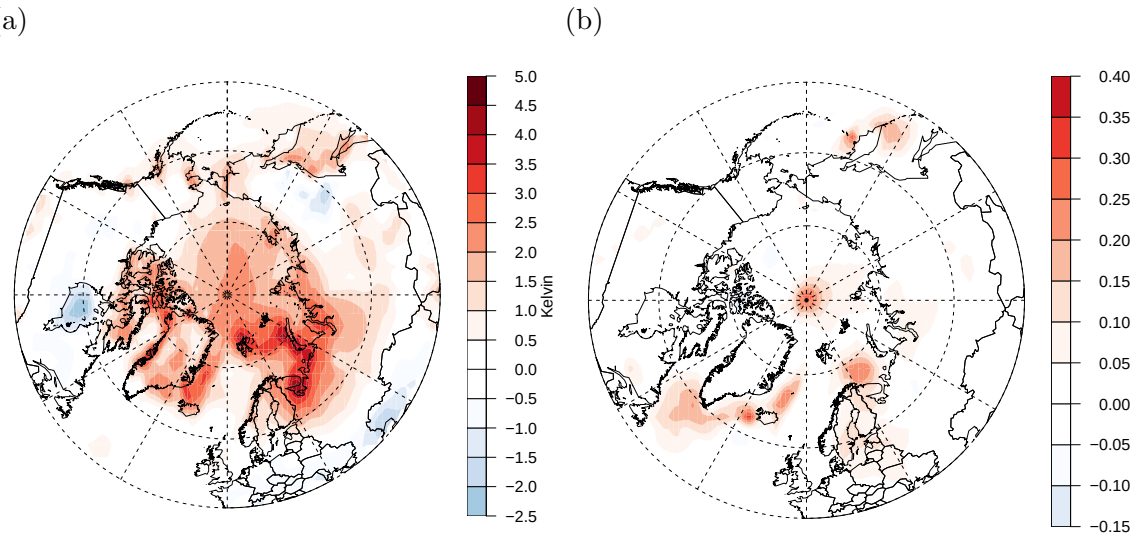

Figure 6.24.: The difference between the posterior mean estimates from the framework developed here and that of Smith et al. (2009), of (a) the historical climate (y_H); and (b) the emergent constraint (λ).

The framework of Tebaldi et al. (2005) and Smith et al. (2009)

The univariate version of the framework described by Smith et al. (2009) and discussed in Section 6.5.2 was also fitted to the surface temperature data analysed in Section 6.8.2. This is identical to the framework of Tebaldi et al. (2005) except that here all the parameters are estimated from the data, whereas the degrees of freedom of the t distribution were previously held fixed. This framework does not allow multiple runs to be included from each model. Therefore, model uncertainty cannot be separated from internal variability. Rather than select one run from each model, the model mean climates $\bar{x}_{sm.}$ were used. This will reduce the impact of internal variability on the estimation of the emergent constraint and should ensure a fair comparison with the posterior uncertainty about the actual climate y_H and climate response y_R. These frameworks also do not allow sampling uncertainty to be separated from measurement error in the observations. Therefore, the combined uncertainty due to internal variability estimated from the models and observation uncertainty estimated from the reanalysis was substituted for σ_z^2 in Equation 6.14, estimated as described in Section 6.8.2, in order to ensure a fair comparison. Vague priors were used for all the parameters. Once again, the burn-in period was found to be very limited, but autocorrelation was extensive. Therefore, the same burn-in and thinning strategies were employed as for the framework in this chapter. The first 20,000 samples were discarded, after which every 2,000th sample was kept until 10,000 samples were obtained from the joint posterior distribution of the parameters.

In Section 6.8.2, the posterior mean estimate of the actual historical climate y_H was several degrees warmer than the expected historical climate of the models μ (Figure 6.14a). The posterior mean estimate y_H from the framework of Smith et al.

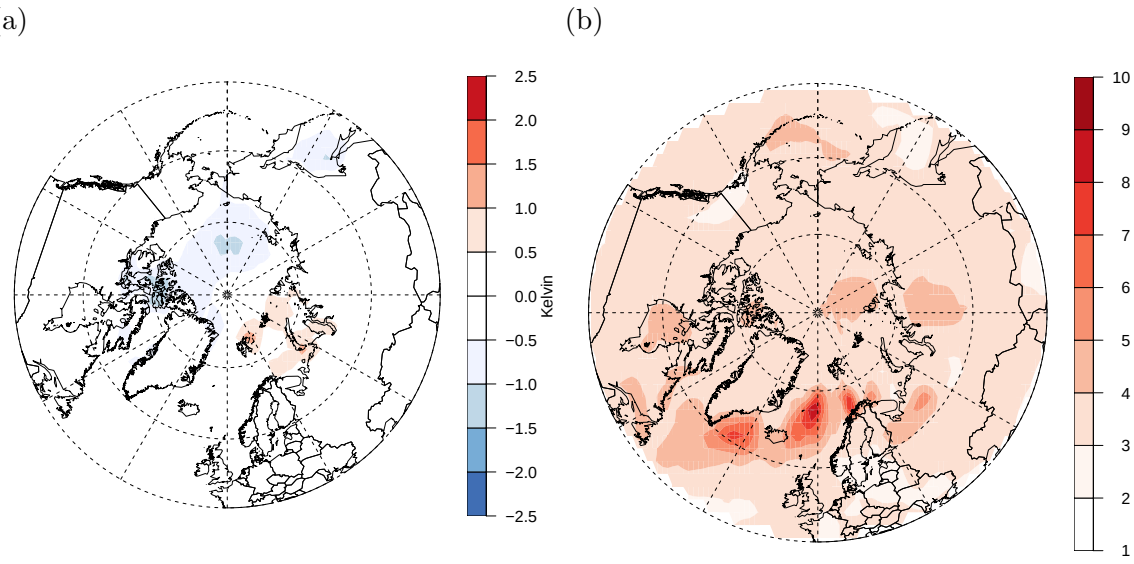

Figure 6.25.: (a) The difference between the posterior means of the projected climate response (y_R) estimated from the framework developed here and that of Smith et al. (2009); and (b) the ratio of the posterior standard error of the climate response (y_R) estimated from the framework developed here and that of Smith et al. (2009).

(2009) is several degrees colder than the estimate from the framework developed here, i.e., closer to the mean climate of the models (Figure 6.24a). This agrees with the findings of Lopez et al. (2006) that the historical climate tended to be heavily influenced by the models in the framework of Tebaldi et al. (2005). The estimate of the emergent constraint is very similar over most of the study area (Figure 6.24b). However, the robustness added by the t distributed model departures does improves the estimates in some regions, notably in the west of the Barents Sea. From Figure 6.21 we noted that FGOALS-g2 had a strong influence on the estimate of the emergent constraint in the exchangeable ensemble. The influence of the outlying model is reduced using the method of Smith et al. (2009), so that the expected negative emergent relationship is estimated.

In the previous section, the projected warming of the actual climate y_R including an emergent constraint was smaller over most of the Arctic than without an emergent constraint(Figure 6.19a). As expected, the estimate from the framework of Smith et al. (2009) lies somewhere between the two estimates over most of the study area (Figure 6.25a). The spatial structure of the difference compared to the framework including an emergent constraint is very similar, but the magnitude is smaller. Finally, the standard error of the projected climate response from the framework developed here is more than three times that estimated by the framework of Smith et al. (2009) over most of the study area (Figure 6.25b). This agrees with the theoretical arguments in Section 6.5.2 and the findings of Lopez et al. (2006) that the uncertainty about the climate response from the framework of Tebaldi et al. (2005) will not span the full range of climate responses simulated by the models.

6.9. Summary

In this chapter, a new Bayesian framework has been developed for combining information from ensembles of climate models and from observations. The proposed framework is based on the concept of a discrepancy between the expected climate of an ensemble of climate models and the actual climate, introduced by Rougier et al. (2013) and Chandler (2013). By comparing those two frameworks we have shown that the "truth plus error" and "exchangeable" approaches will yield identical inferences about the actual climate, provided that the ensemble discrepancy is assumed to arise from a symmetric distribution.

The proposed framework incorporates several important extensions to existing methods. By incorporating all the initial conditions runs available from each climate model, internal variability can be separated from model uncertainty. Chapter 5 showed that this was essential in order to avoid biased estimates of any emergent relationship that might be present. The formulation proposed here shows how emergent relationships can be interpreted as providing constraints on the discrepancy between the expected response of the ensemble, and the actual climate response. This is more flexible than the usual interpretation of emergent constraints, which assumes that the actual climate is jointly exchangeable with the actual climate.

The framework proposed here also separates the effects of observation uncertainty and sampling uncertainty. The process of combining models and observations results in shrinkage of the posterior estimate of the actual historical climate towards the expected historical climate of the models. The shrinkage depends on how informative the models are for the actual climate compared to the uncertainty due to observation and sampling uncertainty. In the presence of an emergent relationship, the posterior estimate of the actual climate response will also experience a shrinkage towards the expected response of the ensemble. The analysis of the cyclone track density data also highlights the fact that our uncertainty about the response of the actual climate due to sampling uncertainty may be almost as great as that due to model discrepancy.

Unfortunately, most observation data does not currently include estimates of the associated uncertainty. Instead, a simple method was proposed and demonstrated for substituting reanalysis data, and estimating our uncertainty about the observed climate from the spread in the reanalyses. Reanalysis data should not be confused with observations. However, it may be the only reasonable source for many climate variables that are indirectly observed, or for which observations are sparse in either space or time. The method proposed integrates easily with the general framework. It should also be applicable to almost any other analysis requiring a ready estimate of the uncertainty associated with the observed climate.

6.10. Discussion

Expressing our beliefs about how informative the models are for the actual climate relative to how informative they are for a new model seems natural. In the analysis presented here, we examined the position of the observations (reanalyses) in the distribution of the modelled climates as a check that our judgements about the prior uncertainty associated with the historical discrepancy were reasonable. If the observations lie in the tails of the distribution of modelled climates at a large number of grid boxes, then we would be forced to conclude that the models were less informative for the actual climate than our prior judgement suggested. Care must be taken when interpreting such results, since this simple test does not account for sampling uncertainty, measurement error, or correlations between grid boxes. However, it is important to ensure that our judgements about the discrepancies are reasonable, or we risk incorporating too much information from the models and obtaining biased inferences for the actual climate.

The comparison with the framework proposed by Rougier et al. (2013) showed that the simpler framework was also compatible with the inclusion of emergent constraints. The idea that the future discrepancy between a climate model and the actual climate might be correlated with the historical discrepancy due to persistence of the historical discrepancy (i.e., bias) was discussed by Rougier (2007). However, the framework presented here represents the first attempt to explicitly interpret emergent constraints (correlations between the *response* and the historical state) in terms of model discrepancy. In fact, the basic form of the joint covariance matrix for the discrepancies proposed by Chandler (2013, Equation 4.2) includes the possibility of persistence but excludes the possibility of an emergent constraint.

The assumption that the expectation of the historical discrepancy is zero is reasonable provided that we have no prior intuition about the sign of the discrepancy (Chandler, 2013). In the case of the Arctic temperature analysis in Section 6.8.2, we have reason to expect a positive discrepancy, since the models are known to underestimate sea ice thickness. Therefore it might have been sensible to incorporate this knowledge into the analysis by specifying a positive value for $\mathrm{E}(\Delta_H)$. Choosing an appropriate value might require a careful elicitation exercise involving one or more experts on the processes involved. Alternatively, if data were already available from another ensemble, then this could be used as the basis for a more informed choice. This might involve specifying additional judgements about the similarity of the two ensembles (e.g., Rougier et al., 2009).

The comparison of the frameworks proposed by Rougier et al. (2013) and Chandler (2013) revealed that identical inferences about the actual climate can be obtained from very different assumptions. This was further demonstrated in the comparison

between the framework developed here and the "truth plus error" approach proposed by Tebaldi et al. (2005). The form of the estimates of the response of the actual climate are very similar, despite the fact that Tebaldi et al. (2005) make no explicit assumption that the emergent constraint should apply to the actual climate response as well as to the model responses. The posterior estimates of the actual climate response differ primarily due to differences in the posterior estimates of the actual historical climate. The assumption by Tebaldi et al. (2005) that the actual climate corresponds to the central tendency of the ensemble leads to more weight being given to the models than the observations when the two are combined (Lopez et al., 2006). This effect would be avoided by the inclusion of an ensemble discrepancy following the generalised "truth plus error" method of Chandler (2013).

The additional robustness imparted by the assumption of t distributed model departures by Tebaldi et al. (2005) proved to be advantageous in the analysis of Arctic temperature. In the Barents Sea, the impact of the heavily biased FGOALS-g2 model was reduced so that the emergent relationship was correctly estimated. The assumption of t distributed departures could be easily incorporated into the hierarchical framework developed in Chapter 5, in order to achieve similar robustness to outlying models.

The results on combining models and observations highlight the role of observation and sampling uncertainty in climate projection, particularly in the presence of emergent relationships. The effect of observation error on projections incorporating emergent constraints has also been recognised by Bhend and Whetton (2013). However, that study considered only observation uncertainty, and assumed that the actual climate was jointly exchangeable ("statistically indistinguishable") with the actual climate. Observation and sampling uncertainty can be difficult to separate. Here we estimated the sampling uncertainty from the climate models. Ideally, this would be estimated from a long time series of observations. Provided that a prior estimate of the observation uncertainty is available, it is still possible to separate the two sources of uncertainty.

One of the innovations introduced in this chapter was the use of reanalysis data in order to estimate the observation uncertainty. For some variables, such as cyclone track density, this may be the only option. Due to the small number of reanalyses available, there may be considerable uncertainty associated with the estimate of their spread. Consequently, the observation uncertainty may be over estimated. However, until more observation data sets include realistic assessments of the associated uncertainties, estimation from reanalyses offers an alternative to ignoring observation uncertainty or relying entirely on prior judgements. Another concern is that reanalyses are also subject to initial condition uncertainty, although the short time steps between data assimilation mean this should be limited. If ensembles of initial

condition members were available for each reanalysis, then this additional source of uncertainty could be estimated separately. However, only the NOAA twentieth century reanalysis (Compo et al., 2011) currently includes multiple members.

7. Conclusion

This chapter summarises the main results contained in this thesis, suggests possible directions for further development, and reflects upon the broader implications for the design and analysis of future multi-model climate change experiments.

7.1. Summary

In Chapter 3, it was shown that a simple two-way ANOVA framework can be used to estimate the expected response of an ensemble of climate models, under the assumption that the models all simulate the same response. The simplifying assumption that all models simulate the same internal variability was proposed in order to estimate the internal variability despite the small initial condition ensembles simulated by each model. Statistical F tests were used to show that if the internal variability is large compared to the differences between the responses simulated by the models, then the models can be assumed to all simulate the same response. If the models all simulate the same response, then it can be argued that there is no reason to expect any discrepancy with the actual response, and the expected response of the ensemble is a good estimate of the actual response.

In Chapter 4, it was argued that multi-model ensembles should be thinned in order to remove the effect of dependence between the outputs of models that share common components. Once thinned to obtain a subset of the models that are judged to be exchangeable, the ensemble can be treated as a random sample from some unknown distribution. A Bayesian hierarchical framework was proposed in order to quantify the structural uncertainty present when the models do not all simulate the same historical climate, or future climate response. The structural uncertainty was separated from the uncertainty due to internal variability using the simplifying assumption proposed in Chapter 3.

In Chapter 5, it was argued that emergent relationships apply only to the differences between the expected climates simulated by the models, and not to departures due to internal variability. It was shown that if the internal variability is large compared to the model uncertainty in the historical scenario, then estimates of emergent con-

straints obtained by simple linear regression will be biased. The proposed Bayesian hierarchical framework was extended to estimate an emergent constraint while accounting for the effect of internal variability. A conditional cross-validation approach was proposed in order to test the robustness of any emergent relationship.

In Chapter 6, the uncertainty about the relationship between the expected climate response of an ensemble of climate models and the actual climate response was represented as a random discrepancy. It was shown that if the discrepancy is assumed to by symmetrically distributed, then the inferences about the actual climate response will be identical, regardless of whether a "truth plus error" or an "exchangeable" approach is used. Emergent constraints were reinterpreted as constraints on the expected value of the discrepancy between the expected response of the ensemble and the actual response. Measurement error in the observations and sampling uncertainty about the actual historical climate were incorporated, and shown to play an important role in the estimation of the constrained response. The idea of an ensemble of reanalyses was introduced, and a simple method for estimating observation uncertainty from reanalysis data was also proposed.

7.2. Directions for further development

The simplifying assumption that climate models all simulate the same internal variability allows internal variability to be separated from model uncertainty. It is difficult to reliably estimate the internal variability simulated by each model individually due to the small number of initial conditions runs available. Alternatively, the internal variability simulated by each model could be estimated by analysing time series rather than relying only on 30-year averages. The simplest approach might be to assume a common linear trend and independent time steps (e.g., Buser et al., 2009; Tebaldi and Sansó, 2009), or a more flexible approach could be taken. The resulting model specific variances could be treated as arising from some common distribution and related to the natural variability of the Earth system through discrepancy terms, analogous to the treatment of the expected climates of the models (similar to the general approach proposed by Chandler, 2013).

The definition of an emergent constraint used in this thesis is perhaps the simplest possible, i.e., linear dependence between the climate change response and the historical state of a single variable. The framework proposed here could be easily extended the estimation of non-linear emergent relationships (as suggested by Bracegirdle and Stephenson, 2012), or to the simultaneous estimation of multiple climate variables (e.g., Tebaldi and Sansó, 2009; Buser et al., 2010). This might include the estimation of correlations and emergent relationships between any or all of the variables

of interest (Karpechko et al., 2013).

The estimation of local effects, such as emergent constraints, could benefit from smoothing over regions larger than the individual grid boxes in order to reduce the risk of biased projections due to the estimation of spurious emergent constraints. Simply aggregating grid boxes is unlikely to be wholly effective, since the spatial pattern of the climate change signal, as well as its magnitude, may vary between models. Furrer et al. (2007b) represented the spatial structure of the climate change response in each model as a combination of spatial basis functions. The coefficients associated with the basis functions for each model were assumed to be drawn from a common multi-variate distribution. This approach would fit well with the general methodology proposed here.

In this thesis, judgements about how informative the climate models are for the actual climate were restricted to prior assessments based on how informative the models are for each other. In principle, a prior distribution could be specified for the variance of the historical discrepancy based on the model uncertainty, and updated using the observations. This approach would also benefit from smoothing over multiple grid boxes. The same spatial basis function approach described above could be adapted to estimate how informative the models are for the historical climate, allowing for variations due to latitude, altitude, land versus ocean, and snow and ice cover, etc. However, this would not explicitly constrain how informative the models are for the actual climate response. It might be possible to assess how informative the models are for the climate response by analysing long paleo-climate simulations. Further research is required on methods to estimate how informative climate models are for the Earth system.

Measurement error combined with sampling uncertainty determines how informative the observations are for the actual climate. Where no estimate of the measurement error is available, a simple method for estimating it from reanalysis data was suggested. This method could easily be applied to multiple observation data sets, but for many variables reanalysis data may be the only option. While observations can be considered independent from climate model output, every reanalysis product has a climate model at its core. Accounting for possible dependences between reanalysis and climate model output is an area for further research, if this multi-reanalysis approach is to be more widely adopted.

7.3. Designing multi-model ensembles

The issue underlying many of the problems addressed in this thesis, in particular model dependence, is really the design of multi-model ensembles. The demand for

probabilistic projections of climate change and the massive computational expense of producing multi-model ensembles mean that this issue is now attracting wider recognition (Katz et al., 2013; Sandgathe et al., 2013). For instance, theoretical attempts have been made to quantify the value of including additional models, (Berliner and Kim, 2008). However, most studies are still limited to the post-hoc interpretation of ensembles of opportunity.

The ensemble thinning approach advocated in this thesis, and by Rougier et al. (2013), suggests possible improvements to the design of future multi-model ensemble experiments. The importance of accounting for internal variability in model projections has been demonstrated repeatedly throughout this thesis. Daron and Stainforth (2013) suggest that an even wider range of initial conditions should be explored. If each centre submitted results from only one model, then the computational resources previously allocated to model variants could be used to perform additional initial condition runs. Alternatively, each centre could nominate a canonical model for each group of experiments. Researchers could then be encouraged to analyse only the outputs from the canonical models, in order to reduce the impact of model dependence.

Ultimately, the design of multi-model ensembles is limited by the difficulty of defining a model space from which different model designs could be sampled systematically. Given the fact that large perturbed physics ensembles can yield a larger spread of outcomes than multi-model ensembles (Murphy et al., 2004; Stainforth et al., 2005), should the role of the multi-model ensembles be re-evaluated? There are specific questions that could be answered using properly designed multi-model ensembles, e.g., the uncertainty due to the choice of grid from a well defined class of grids. Using statistical emulation techniques (e.g., O'Hagan, 2006; Rougier et al., 2009; Sexton et al., 2012; Williamson et al., 2013), it might be possible to combine results from designed multi-model ensembles with perturbed physics ensembles, in order to better quantify the effects of structural uncertainty.

7.4. Adoption of statistical methods for climate projection

The adoption of statistical methods for making inferences about future climate change has been slow. It is to be hoped that the development of frameworks that address the issues of model dependence and model inadequacy, such as the one proposed in this thesis, will be welcomed by the climate science community. However, the perceived complexity of such frameworks may still present a barrier to acceptance. Even when the computer code required to perform the necessary es-

timation has been made available (e.g., Tebaldi et al., 2005; Smith et al., 2009), adoption has not been widespread. Chandler (2013) derived explicit expressions for the posterior mean and variance of the actual climate, under the assumption that all components were normally distributed with known variances. Rougier et al. (2013) achieved the same goal using the second order Bayes Linear methodology proposed by Goldstein and Wooff (2007). Whether or not these simpler analytic solutions are more readily accepted remains to be seen. The potential for statistical methods to quantify and reduce uncertainty about projections of future climate has been clearly demonstrated. However, continued and closer co-operation between statisticians and climate scientists is required in order to fully realise the benefits.

7.5. Conclusion

This thesis has proposed new frameworks for making inferences about future climate change based on the outputs of an ensemble of climate models, and observations of the recent climate. The most sophisticated framework is able to separate structural uncertainty from internal variability, provide unbiased estimates of emergent constraints, quantify the contributions of both sampling uncertainty and measurement error, and account for both model dependence and model inadequacy.

Appendices

A. Background to the analysis of variance frameworks

A.1. Derivation of the two-way framework with interactions

The log-likelihood of the two-way framework with interactions (Eqn. 3.4) is

$$\ell\left(\boldsymbol{\beta};\mathbf{x}\right) \propto -\frac{N_{\cdot\cdot}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{N_{sm}}\left(x_{smr}-\mu-\alpha_m-\beta_s-\gamma_{sm}\right)^2$$

Maximum likelihood estimates of the parameters are obtained by maximising the log-likelihood with respect to all the parameters simultaneously, subject to the constraints that $\sum_{m=1}^{M}\alpha_m = 0$, $\beta_H = 0$, $\gamma_{Hm} = 0\ \forall\ m = 1,\ldots,M$ and $\sum_{m=1}^{M}\gamma_{Fm} = 0$. This is equivalent to taking the partial derivative of the log likelihood with respect to each parameter, setting each equation equal to zero, and solving the resulting set of simultaneous equations with the help of Lagrange multipliers to ensure that the constraints are met. Solving those equations yields the following estimates

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\overline{x}_{Hm\cdot} \tag{A.1a}$$

$$\hat{\alpha}_m = \overline{x}_{Hm\cdot} - \hat{\mu} \quad \forall\ m = 1,\ldots,M \tag{A.1b}$$

$$\hat{\beta}_F = \frac{1}{M}\sum_{m=1}^{M}\overline{x}_{Fm\cdot} - \frac{1}{M}\sum_{m=1}^{M}\overline{x}_{Hm\cdot} \tag{A.1c}$$

$$\hat{\gamma}_{Fm} = \overline{x}_{Fm\cdot} - \overline{x}_{Hm\cdot} - \hat{\beta}_F \quad \forall\ m = 1,\ldots,M \tag{A.1d}$$

and

$$s^2 = \hat{\sigma}^2 = \frac{1}{N_{\cdot\cdot}}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{N_{sm}}\left(x_{smr}-\hat{x}_{smr}\right)^2$$

however, the maximum likelihood estimate of σ^2 is known to be biased (Davison, 2003), an unbiased estimate is given by

$$s^2 = \hat{\sigma}^2 = \frac{1}{N_{\cdot\cdot} - P} \sum_{m=1}^{M} \sum_{s \in \{H,F\}} \sum_{r=1}^{N_{sm}} (x_{smr} - \hat{x}_{smr})^2 \tag{A.2}$$

where $P = 2M$ is the number of mean parameters to be estimated and $\hat{x}_{smr}$ are the fitted values given by

$$\hat{x}_{smr} = \hat{\mu} + \hat{\alpha}_m + \hat{\beta}_s + \hat{\gamma}_{sm} = \overline{x}_{sm\cdot} \tag{A.3}$$

The sampling variances of the parameter estimates are derived by taking the variance of the maximum likelihood estimates in Equation A.1

$$\text{var}\left(\hat{\mu}\right) = \frac{\sigma^2}{M^2} \sum_{m=1}^{M} \frac{1}{N_{Hm}} \tag{A.4a}$$

$$\text{var}\left(\hat{\alpha}_m\right) = \frac{\sigma^2}{N_{Hm}} \left(\frac{M-2}{M}\right) + \text{var}\left(\hat{\mu}\right) \quad \forall\ m = 1, \ldots, M \tag{A.4b}$$

$$\text{var}\left(\hat{\beta}_F\right) = \frac{\sigma^2}{M^2} \sum_{m=1}^{M} \frac{N_{\cdot m}}{N_{Hm} N_{Fm}} \tag{A.4c}$$

$$\text{var}\left(\hat{\gamma}_{Fm}\right) = \sigma^2 \left(\frac{M-2}{M}\right) \frac{N_{\cdot m}}{N_{Hm} N_{Fm}} + \text{var}\left(\hat{\beta}_F\right) \quad \forall\ m = 1, \ldots, M \tag{A.4d}$$

and

$$\text{var}\left(\hat{x}_{smr}\right) = \frac{\sigma^2}{N_{sm}} \tag{A.5}$$

The residuals are

$$e_{smr} = x_{smr} - \hat{x}_{smr} = x_{smr} - \overline{x}_{sm\cdot} \tag{A.6}$$

and their sampling variance is

$$\text{var}\left(e_{smr}\right) = \sigma^2 \left(1 - \frac{1}{N_{sm}}\right) \tag{A.7}$$

so the standardised residuals are given by

$$e'_{smr} = \frac{e_{smr}}{\sqrt{\text{var}\left(e_{smr}\right)}} \tag{A.8}$$

A.2. Derivation of the two-way framework

The log-likelihood of the two-way framework (Equation 3.5) is:

$$\ell\left(\boldsymbol{\beta};\mathbf{x}\right) \propto -\frac{N_{\cdot\cdot}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{N_{sm}}\left(x_{smr}-\mu-\alpha_m-\beta_s\right)^2$$

subject to the constraints that $\sum_{m=1}^{M}\alpha_m = 0$, $\beta_H = 0$. The maximum likelihood estimates of the parameters are derived in the same way as for the framework with interactions in the previous section

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\left(\overline{x}_{\cdot m\cdot} - \frac{N_{Fm}}{N_{\cdot m}}\hat{\beta}_F\right) \tag{A.9a}$$

$$\hat{\alpha}_m = \overline{x}_{\cdot m\cdot} - \frac{N_{Fm}}{N_{\cdot m}}\hat{\beta}_F - \hat{\mu} \quad \forall\ m = 1,\ldots,M \tag{A.9b}$$

$$\hat{\beta}_F = \frac{1}{\sum_{m=1}^{M} w_m}\sum_{m=1}^{M} w_m\left(\overline{x}_{Fm\cdot} - \overline{x}_{Hm\cdot}\right) \tag{A.9c}$$

where

$$\overline{x}_{\cdot m\cdot} = \sum_{s\in\{H,F\}}\sum_{r=1}^{N_{sm}} x_{smr} \quad \text{and} \quad w_m = \frac{N_{Hm}N_{Fm}}{N_{\cdot m}}$$

Once again, σ^2 is also unknown and must be estimated, so it is replaced by the estimate s^2 from Equation A.2 with $P = M + 1$ and fitted values $\hat{x}_{smr}$ given by

$$\hat{x}_{smr} = \begin{cases} \overline{x}_{\cdot m\cdot} - \frac{N_{Fm}}{N_{\cdot m}}\hat{\beta}_F & \text{if } s = H, \\ \overline{x}_{\cdot m\cdot} + \frac{N_{Hm}}{N_{\cdot m}}\hat{\beta}_F & \text{if } s = F. \end{cases} \tag{A.10}$$

The sampling variances of the parameter estimates are derived by taking the variance of the the maximum likelihood estimates in Equation A.9

$$\operatorname{var}\left(\hat{\mu}\right) = \frac{\sigma^2}{M^2}\left(\frac{1}{w_{\cdot}}\left(\sum_{m=1}^{M}\frac{N_{Fm}}{N_{\cdot m}}\right)^2 + \sum_{m=1}^{M}\frac{1}{N_{\cdot m}}\right) \tag{A.11a}$$

$$\operatorname{var}\left(\hat{\alpha}_m\right) = \frac{\sigma^2}{N_{\cdot m}}\left(\frac{M-2}{M}\right) + \frac{\sigma^2}{M^2}\sum_{m=1}^{M}\frac{1}{N_{\cdot m}} + \frac{\sigma^2}{w_{\cdot}}\left(\frac{N_{Fm}}{N_{\cdot m}} - \frac{1}{M}\sum_{m=1}^{M}\frac{N_{Fm}}{N_{\cdot m}}\right)^2 \tag{A.11b}$$

$$\operatorname{var}\left(\hat{\beta}_F\right) = \frac{\sigma^2}{w_{\cdot}} \tag{A.11c}$$

where $w_{\cdot} = \sum_{m=1}^{M} w_m$ and

$$\operatorname{var}(\hat{x}_{smr}) = \begin{cases} \frac{\sigma^2}{N_{\cdot m}} + \frac{\sigma^2}{w_{\cdot}} \left(\frac{N_{Fm}}{N_{\cdot m}}\right)^2 & \text{if } s = H, \\ \frac{\sigma^2}{N_{\cdot m}} + \frac{\sigma^2}{w_{\cdot}} \left(\frac{N_{Hm}}{N_{\cdot m}}\right)^2 & \text{if } s = F. \end{cases} \tag{A.12}$$

The residuals are given by

$$e_{smr} = x_{smr} - \hat{x}_{smr} = \begin{cases} x_{Hmr} - \overline{x}_{\cdot m \cdot} + \frac{N_{Fm}}{N_{\cdot m}} \hat{\beta}_F & \text{if } s = H, \\ x_{Fmr} - \overline{x}_{\cdot m \cdot} - \frac{N_{Hm}}{N_{\cdot m}} \hat{\beta}_F & \text{if } s = F. \end{cases} \tag{A.13}$$

and their sampling variances are

$$\operatorname{var}(e_{smr}) = \begin{cases} \sigma^2 - \frac{\sigma^2}{N_{\cdot m}} - \frac{\sigma^2}{w_{\cdot}} \left(\frac{N_{Fm}}{N_{\cdot m}}\right)^2 & \text{if } s = H, \\ \sigma^2 - \frac{\sigma^2}{N_{\cdot m}} - \frac{\sigma^2}{w_{\cdot}} \left(\frac{N_{Hm}}{N_{\cdot m}}\right)^2 & \text{if } s = F. \end{cases} \tag{A.14}$$

and the standardised residuals are given be Equation A.8, substituting from Equations A.13 and A.14.

A.3. Derivation of the one-way framework

The log-likelihood of the one-way framework (Equation 3.7) is

$$\ell(\boldsymbol{\beta}; \mathbf{x}) \propto -\frac{N_{\cdot\cdot}}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{m=1}^{M} \sum_{s \in \{H,F\}} \sum_{r=1}^{N_{sm}} (x_{smr} - \mu - \beta_s)^2$$

with the constraint that $\beta_H = 0$. The maximum likelihood estimates of the parameters are derived in the same way as the two-way frameworks

$$\hat{\mu} = \frac{1}{N_{H\cdot}} \sum_{m=1}^{M} N_{Hm} \overline{x}_{Hm\cdot} \tag{A.15a}$$

$$\hat{\beta_F} = \frac{1}{N_{F\cdot}} \sum_{m=1}^{M} N_{Fm} \overline{x}_{Fm\cdot} - \frac{1}{N_{H\cdot}} \sum_{m=1}^{M} N_{Hm} \overline{x}_{Hm\cdot} \tag{A.15b}$$

$$\tag{A.15c}$$

The internal variability σ^2 is also unknown and must be estimated, so it is replaced by the estimate s^2 from Equation A.2 with $P = 2$ and fitted values $\hat{x}_{smr}$ given by

$$\hat{x}_{smr} = \hat{\mu} + \hat{\beta}_s = \sum_{m=1}^{M} \sum_{r=1}^{N_{sm}} \frac{x_{smr}}{N_{s\cdot}} = \bar{x}_{s\cdot\cdot} \tag{A.16}$$

The sampling variances of the parameter estimates are obtained by taking the variance of the maximum likelihood estimates in Equation A.15

$$\mathrm{var}\left(\hat{\mu}\right) = \frac{\sigma^2}{N_{H.}} \tag{A.17a}$$

$$\mathrm{var}\left(\hat{\beta}_F\right) = \frac{\sigma^2}{N_{F.}} + \frac{\sigma^2}{N_{H.}} \tag{A.17b}$$

and

$$\mathrm{var}\left(\hat{x}_{smr}\right) = \mathrm{var}\left(\overline{x}_{s..}\right) = \frac{\sigma^2}{N_{s.}} \tag{A.18}$$

The residuals are

$$e_{smr} = x_{smr} - \hat{x}_{smr} = x_{smr} - \sum_{m=1}^{M}\sum_{r=1}^{N_{sm}} x_{smr} \tag{A.19}$$

and their sampling variance is

$$\mathrm{var}\left(e_{smr}\right) = \sigma^2\left(1 - \frac{1}{N_{s.}}\right) \tag{A.20}$$

and the standardised residuals are given be Equation A.8, substituting from Equations A.19 and A.20.

A.4. Estimator biases

If the ensemble expected climate response β_F is estimated using the two-way framework but the "true" framework includes model specific response departures γ_{Fm}, then the bias of the estimate is given by

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_F - \beta_F\right) &= \mathrm{E}\left(\frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\right)\right) - \beta_F \\
&= \frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\,\mathrm{E}\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\right) - \beta_F \\
&= \frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\left(\left[\mu + \alpha_m + \beta_F + \gamma_{Fm}\right] - \left[\mu + \alpha_m\right]\right) - \beta_F \\
&= \frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\gamma_{Fm}
\end{aligned} \tag{A.21}$$

If the ensemble expected climate response β_F is estimated using the one-way framework but the "true" framework includes model specific historical and response de-

partures α_m and γ_{Fm}, then the bias of the estimate is given by

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_F - \beta_F\right) &= \mathrm{E}\left(\frac{1}{N_{F.}}\sum_{m=1}^{M} N_{Fm}\overline{x}_{Fm.} - \frac{1}{N_{H.}}\sum_{m=1}^{M} N_{Hm}\overline{x}_{Hm.}\right) - \beta_F \\
&= \frac{1}{N_{F.}}\sum_{m=1}^{M} N_{Fm}\,\mathrm{E}\left(\overline{x}_{Fm.}\right) - \frac{1}{N_{H.}}\sum_{m=1}^{M} N_{Hm}\,\mathrm{E}\left(\overline{x}_{Hm.}\right) - \beta_F \\
&= \frac{1}{N_{F.}}\sum_{m=1}^{M} N_{Fm}\left[\mu + \alpha_m + \beta_F + \gamma_{Fm}\right] - \frac{1}{N_{H.}}\sum_{m=1}^{M} N_{Hm}\left[\mu + \alpha_m\right] - \beta_F \\
&= \frac{1}{N_{F.}}\sum_{m=1}^{M} N_{Fm}\alpha_m - \frac{1}{N_{H.}}\sum_{m=1}^{M} N_{Hm}\alpha_m + \frac{1}{N_{F.}}\sum_{m=1}^{M} N_{Fm}\gamma_{Fm} \qquad \text{(A.22)}
\end{aligned}$$

If the ensemble expected climate response β_F is estimated using the framework with interactions but the "true" framework does not include any model specific departures α_m or γ_{Fm}, then the bias of the estimate is given by

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_F - \beta_F\right) &= \mathrm{E}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\right)\right) - \beta_F \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathrm{E}\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\right) - \beta_F \\
&= \frac{1}{M}\sum_{m=1}^{M}\left(\left[\mu + \beta_F\right] - \left[\mu\right]\right) - \beta_F \\
&= 0 \qquad \text{(A.23)}
\end{aligned}$$

so the estimate from the framework with interactions is unbiased even if the models all simulate the same historical climate and climate response.

If the ensemble expected climate response β_F is estimated using the two-way framework but the "true" framework does not include any model specific departures α_m or γ_{Fm}, then the bias of the estimate is given by

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_F - \beta_F\right) &= \mathrm{E}\left(\frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\right)\right) - \beta_F \\
&= \frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\,\mathrm{E}\left(\overline{x}_{Fm.} - \overline{x}_{Hm.}\right) - \beta_F \\
&= \frac{1}{w_{\cdot}}\sum_{m=1}^{M} w_m\left(\left[\mu + \beta_F\right] - \left[\mu\right]\right) - \beta_F \\
&= 0 \qquad \text{(A.24)}
\end{aligned}$$

so the estimate from the two-way framework is unbiased even if the models all simulate the same historical climate.

A.5. The relationship between f^2 and Ψ

Let F be a random variable with non-central F distribution with ν_1 and ν_2 degrees of freedom and non-centrality parameter λ. The expectation of F is

$$\mathrm{E}(F) = \frac{\nu_2(\nu_1 + \lambda)}{\nu_1(\nu_2 - 2)} \tag{A.25}$$

Then for $F = \frac{\nu_2}{\nu_1} f^2$ with fixed degrees of freedom $\nu 1$ and $\nu 2$

$$\mathrm{E}\left(f^2\right) = \frac{\nu_1 + \lambda}{\nu_2 - 2}$$

Then assuming a balanced ensemble, i.e., $N_{sm} = N \; \forall \; s, m$, we have

$$\lambda_\gamma = N(M-1)\Psi_\gamma^2/2$$

from Equation 3.11, with $\nu_1 = M - 1$ and $\nu_2 = N_{\cdot\cdot} - 2M = 2MN - 2M$ so that

$$\mathrm{E}\left(f_\gamma^2\right) = \frac{2(M-1) + N(M-1)\Psi_\gamma^2}{2(2MN - 2M - 2)}$$

and in the limit as $M \to \infty$

$$\mathrm{E}\left(f_\gamma^2\right) = \frac{M-1}{M} \frac{2 + N\Psi_\gamma^2}{4(N-1) - 4/M} \approx \frac{2 + N\Psi_\gamma^2}{4(N-1)}$$

and in the limit as $N \to \infty$

$$\mathrm{E}\left(f_\gamma^2\right) \approx \frac{1}{2(N-1)} + \frac{N}{N-1}\frac{\Psi_\gamma^2}{4} \approx \frac{\Psi_\gamma^2}{4} \tag{A.26}$$

and from Equation 3.16 we have

$$\lambda_\alpha = 2N(M-1)\Psi_\alpha^2$$

with $\nu_1 = M - 1$ and $\nu_2 = N_{\cdot\cdot} - 2 = 2MN - 2$ so that

$$\mathrm{E}\left(f_\alpha^2\right) = \frac{(M-1) + 2N(M-1)\Psi_\alpha^2}{(2MN - 2) - 2}$$

and in the limit as $M \to \infty$

$$\mathrm{E}\left(f_\alpha^2\right) = \frac{M-1}{M} \frac{1 + 2N\Psi_\alpha^2}{2(N - 2/M)} \approx \frac{1 + 2N\Psi_\alpha^2}{2N}$$

and in the limit as $N \to \infty$

$$\mathrm{E}\left(f_\alpha^2\right) \approx \frac{1}{2N} + \Psi_\alpha^2 \approx \Psi_\alpha^2 \tag{A.27}$$

B. Background to the hierarchical framework

B.1. Derivation of the full conditional distributions

The ensemble components are assumed to have the following distributions (Equation 4.3)

$$\begin{aligned} x_{Hmr} &\overset{iid}{\sim} N\left(\mu + \alpha_m \ , \ \tau_H^{-1}\right) \\ x_{Fmr} &\overset{iid}{\sim} N\left(\mu + \alpha_m + \beta + \gamma_m \ , \ \tau_F^{-1}\right) \\ \alpha_m &\overset{iid}{\sim} N\left(0 \ , \ \tau_\alpha^{-1}\right) \\ \gamma_m &\overset{iid}{\sim} N\left(0 \ , \ \tau_\gamma^{-1}\right) \end{aligned}$$

with the following prior distributions for the mean parameters (Equation 4.2

$$\begin{aligned} \mu &\sim N\left(a_\mu \ , \ b_\mu^{-1}\right) \\ \beta &\sim N\left(a_\beta \ , \ b_\beta^{-1}\right) \end{aligned}$$

and the following priors for the precision parameters (Equation 4.4)

$$\begin{aligned} \tau_H &\sim Gamma\left(c_H \ , \ d_H\right) \\ \tau_F &\sim Gamma\left(c_F \ , \ d_F\right) \\ \tau_\alpha &\sim Gamma\left(c_\alpha \ , \ d_\alpha\right) \\ \tau_\gamma &\sim Gamma\left(c_\gamma \ , \ d_\gamma\right) \end{aligned}$$

The likelihood of the hierarchical model is

$$\begin{aligned} \Pr\left(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}\right) \propto \tau_H^{N_{H.}/2} \exp\left(-\frac{\tau_H}{2} \sum_{m=1}^{M} \sum_{r=1}^{N_{Hm}} \left(x_{Hmr} - \mu - \alpha_m\right)^2\right) \\ \tau_F^{N_{F.}/2} \exp\left(-\frac{\tau_F}{2} \sum_{m=1}^{M} \sum_{r=1}^{N_{Fm}} \left(x_{Fmr} - \mu - \alpha_m - \beta - \gamma_m\right)^2\right) \end{aligned} \tag{B.1}$$

where $\mathbf{x} = (x_{smr} \ \forall \ s, m, r)$ are the model runs, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_M, \gamma_1, \ldots, \gamma_M)$ is the vector of random effects, and $\boldsymbol{\phi} = \left(\mu, \beta, \sigma_H^2, \sigma_F^2, \sigma_\alpha^2, \sigma_\gamma^2\right)$ is the vector of parameters. The prior probability of the random effects, conditional on the parameters is

$$\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\phi}\right) \propto \tau_\alpha^{M/2} \exp\left(-\frac{\tau_\alpha}{2} \sum_{m=1}^{M} \left(\alpha_m - 0\right)^2\right) \tau_\gamma^{M/2} \exp\left(-\frac{\tau_\gamma}{2} \sum_{m=1}^{M} \left(\gamma_m - 0\right)^2\right) \quad \text{(B.2)}$$

and the prior probabilities of the parameters are

$$\begin{aligned}\Pr\left(\boldsymbol{\phi}\right) \propto & \exp\left(-\frac{b_\mu}{2}\left(\mu - a_\mu\right)^2\right) \exp\left(-\frac{b_\beta}{2}\left(\beta - a_\beta\right)^2\right) \\ & \tau_H^{c_H - 1} \exp\left(-d_H \tau_H\right) \tau_F^{c_F - 1} \exp\left(-d_F \tau_F\right) \tau_\alpha^{c_\alpha - 1} \exp\left(-d_\alpha \tau_\alpha\right) \tau_\gamma^{c_\gamma - 1} \exp\left(-d_\gamma \tau_\gamma\right)\end{aligned} \quad \text{(B.3)}$$

so that the joint posterior of the parameters is given by

$$\Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right) \propto \Pr\left(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi}\right)$$

up to a constant of integration. The full conditional distributions of the parameters are found by selecting the terms in $\Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right)$ which include a particular parameter and simplifying. For example

$$\begin{aligned}\Pr\left(\mu | \ldots\right) \propto & \exp\left(-\frac{\tau_H}{2} \sum_{m=1}^{M} \sum_{r=1}^{N_{Hm}} \left(x_{Hmr} - \mu - \alpha_m\right)^2 - \right. \\ & \left. \frac{\tau_F}{2} \sum_{m=1}^{M} \sum_{r=1}^{N_{Fm}} \left(x_{Fmr} - \mu - \alpha_m - \beta - \gamma_m\right)^2 - \frac{b_\mu}{2}\left(\mu - a_\mu\right)^2\right) \\ \propto & \exp\left(-\frac{\tau_H}{2} \sum_{m=1}^{M} \sum_{r=1}^{N_{Hm}} \left(\mu^2 - 2\left(x_{Hmr} - \alpha_m\right)\right)\mu - \right. \\ & \left. \frac{\tau_F}{2} \sum_{m=1}^{M} \sum_{r=1}^{N_{Fm}} \left(\mu^2 - 2\left(x_{Fmr} - \alpha_m - \beta - \gamma_m\right)\mu\right) - \frac{b_\mu}{2}\left(\mu^2 - 2a_\mu\mu\right)\right) \\ \propto & \exp\left(-\frac{1}{2}\left[\left(N_{H.}\tau_H + N_{F.}\tau_F + b_\mu\right)\mu^2 \right.\right. \\ & \left.\left. -2\left(\tau_H \sum_{m=1}^{M} N_{Hm}\left(\overline{x}_{Hm.} - \alpha_m\right) + \tau_F \sum_{m=1}^{M} N_{Fm}\left(\overline{x}_{Fm.} - \alpha_m - \beta - \gamma_m\right) + b_\mu a_\mu\right)\mu\right]\right)\end{aligned}$$

which we recognise as having the quadratic form of a normal distribution, so that

$$\begin{aligned}\mu | \ldots \sim N\left(\right. & \frac{b_\mu a_\mu + \tau_H \sum_{m=1}^{M} N_{Hm}\left(\overline{x}_{Hm.} - \alpha_m\right) + \tau_F \sum_{m=1}^{M} N_{Fm}\left(\overline{x}_{Fm.} - \alpha_m - \beta - \gamma_m\right)}{b_\mu + N_{H.}\tau_H + N_{F.}\tau_F}, \\ & \left.\left(b_\mu + N_{H.}\tau_H + N_{F.}\tau_F\right)^{-1}\right)\end{aligned} \quad \text{(B.4)}$$

The posterior distributions of the remaining parameters are derived in the same way,

so that

$$\beta|\ldots \sim N\left(\frac{b_\beta a_\beta + \tau_F \sum_{m=1}^{M} N_{Fm}\left(\overline{x}_{Fm.} - \mu - \alpha_m - \gamma_m\right)}{b_\beta + N_{F.}\tau_F}\ ,\ \left(b_\beta + N_{F.}\tau_F\right)^{-1}\right) \tag{B.5}$$

$$\alpha_m|\ldots \sim N\left(\frac{\tau_H N_{Hm}\left(\overline{x}_{Hm.} - \mu\right) + \tau_F N_{Fm}\left(\overline{x}_{Fm.} - \mu - \beta - \gamma_m\right)}{N_{Hm}\tau_H + N_{Fm}\tau_F + \tau_\alpha}\ ,\ \left(N_{Hm}\tau_H + N_{Fm}\tau_F + \tau_\alpha\right)^{-1}\right) \tag{B.6}$$

$$\gamma_m|\ldots \sim N\left(\frac{\tau_F N_{Fm}\left(\overline{x}_{Fm.} - \mu - \alpha_m - \beta\right)}{N_{Fm}\tau_F + \tau_\gamma}\ ,\ \left(N_{Fm}\tau_F + \tau_\gamma\right)^{-1}\right) \tag{B.7}$$

$$\tau_H|\ldots \sim Gamma\left(c_H + \frac{N_{H.}}{2}\ ,\ d_H + \frac{\sum_{m=1}^{M}\sum_{r=1}^{N_{Hm}}\left(x_{Hmr} - \mu - \alpha_m\right)^2}{2}\right) \tag{B.8}$$

$$\tau_F|\ldots \sim Gamma\left(c_F + \frac{N_{F.}}{2}\ ,\ d_F + \frac{\sum_{m=1}^{M}\sum_{r=1}^{N_{Fm}}\left(x_{Fmr} - \mu - \alpha_m - \beta - \gamma_m\right)^2}{2}\right) \tag{B.9}$$

$$\tau_\alpha|\ldots \sim Gamma\left(c_\alpha + \frac{M}{2}\ ,\ d_\alpha + \frac{\sum_{m=1}^{M}\left(\alpha_m - 0\right)^2}{2}\right) \tag{B.10}$$

$$\tau_\gamma|\ldots \sim Gamma\left(c_\gamma + \frac{M}{2}\ ,\ d_\gamma + \frac{\sum_{m=1}^{M}\left(\gamma_m - 0\right)^2}{2}\right) \tag{B.11}$$

B.2. Equivalence of cross-validation methods

Suppose that we wish to predict some outcome $\tilde{x}$ of running a new model j, then the predictive distribution is

$$\Pr\left(\tilde{x} \mid \mathbf{x}\right) = \int_{\boldsymbol{\phi}} \Pr\left(\tilde{x} \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}\right) \Pr\left(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi} \mid \mathbf{x}\right) d\boldsymbol{\phi} \tag{B.12}$$

The p-value required for cross-validation is $\Pr\left(\tilde{x} > x^o \mid \mathbf{x}\right)$, where x^o is the observed

output of the model. However, it is more convenient to consider the alternative p-value $\Pr(\tilde{x} \leq x^o \mid \mathbf{x}) = 1 - \Pr(\tilde{x} > x^o \mid \mathbf{x})$, which is given by

$$\Pr(\tilde{x} \leq x^o \mid \mathbf{x}) = \int_{\tilde{x}=-\infty}^{\tilde{x}=x^o} \int_{\boldsymbol{\phi}} \Pr\left(\tilde{x} \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}\right) \Pr\left(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}\right) \Pr(\boldsymbol{\phi} \mid \mathbf{x}) \, d\boldsymbol{\phi} d\tilde{x} \tag{B.13}$$

$$= \int_{\boldsymbol{\phi}} \Pr\left(\tilde{x} \leq x^o \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}\right) \Pr\left(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\phi}\right) \Pr(\boldsymbol{\phi} \mid \mathbf{x}) \, d\boldsymbol{\phi} \tag{B.14}$$

Note that in Smith et al. (2009), the probability integral transformation U is actually $U = \Pr\left(\tilde{x} \leq x^o \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}\right)$. So the cross validation method described in Section 4.6.3 is equivalent to Equation B.13 and the method of Smith et al. (2009) is equivalent to Equation B.14. In either case, we do not know the posterior distribution of the parameters $\Pr(\boldsymbol{\phi} \mid \mathbf{x})$, we only have N samples from it. So Equation B.13 is approximated by computing N samples of $\tilde{x}$ from $\Pr\left(\tilde{x} \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}\right)$, one for each sample from $\Pr(\boldsymbol{\phi} \mid \mathbf{x})$, and then calculating

$$\Pr(\tilde{x} \leq x^o \mid \mathbf{x}) \approx \frac{1}{N} \sum_{n=1}^{N} \mathrm{I}\left(\tilde{x}^{(n)} \leq x^o\right) \tag{B.15}$$

and Equation B.14 is approximated by computing $\Pr\left(\tilde{x} \leq x^o \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}^{(n)}\right)$ for each of the N samples from $\Pr(\boldsymbol{\phi} \mid \mathbf{x})$, and then calculating

$$\Pr(\tilde{x} \leq x^o \mid \mathbf{x}) \approx \frac{1}{N} \sum_{n=1}^{N} \Pr\left(\tilde{x} \leq x^o \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}^{(n)}\right) \tag{B.16}$$

which is equivalent to Equation B.15 since $\tilde{x}^{(n)} \leq x^o$ with probability $\Pr\left(\tilde{x} \leq x^o \mid \tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}^{(n)}\right)$.

C. Background to the extended hierarchical framework

C.1. Derivation of the full conditional distributions

The ensemble components are assumed to have the following distributions

$$\begin{aligned} x_{Hmr} &\overset{iid}{\sim} N\left(\mu + \alpha_m \ , \ \tau_H^{-1}\right) \\ x_{Fmr} &\overset{iid}{\sim} N\left(\mu + \alpha_m + \beta + \gamma_m \ , \ \tau_F^{-1}\right) \\ \alpha_m &\overset{iid}{\sim} N\left(0 \ , \ \tau_\alpha^{-1}\right) \\ \gamma_m \mid \alpha_m &\overset{iid}{\sim} N\left(\lambda\alpha_m \ , \ \tau_{\gamma|\alpha}^{-1}\right) \end{aligned}$$

with the following prior distributions

$$\begin{aligned} \mu &\sim N\left(a_\mu \ , \ b_\mu^{-1}\right) \\ \beta &\sim N\left(a_\beta \ , \ b_\beta^{-1}\right) \\ \lambda &\sim N\left(a_\lambda \ , \ b_\lambda^{-1}\right) \\ \tau_H &\sim Gamma\left(c_H \ , \ d_H\right) \\ \tau_F &\sim Gamma\left(c_F \ , \ d_F\right) \\ \tau_\alpha &\sim Gamma\left(c_\alpha \ , \ d_\alpha\right) \\ \tau_{\gamma|\alpha} &\sim Gamma\left(c_{\gamma|\alpha} \ , \ d_{\gamma|\alpha}\right) \end{aligned}$$

where $a_\mu = a_\beta = a_\lambda = 0$ and $b_\mu = b_\beta = b_\lambda = 10^{-6}$ and $c_H = c_F = c_\alpha = c_{\gamma|\alpha} = d_H = d_F = 10^{-3}$ and $d_\alpha = d_{\gamma|\alpha} = 10^{-1}$. The likelihood of the extended hierarchical model is given by

$$\begin{aligned} \Pr\left(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}\right) \propto \tau_H^{N_{H.}/2} \exp\left(-\frac{\tau_H}{2}\sum_{m=1}^{M}\sum_{r=1}^{N_{Hm}}\left(x_{Hmr} - \mu - \alpha_m\right)^2\right) \\ \tau_F^{N_{F.}/2} \exp\left(-\frac{\tau_F}{2}\sum_{m=1}^{M}\sum_{r=1}^{N_{Fm}}\left(x_{Fmr} - \mu - \alpha_m - \beta - \gamma_m\right)^2\right) \end{aligned} \tag{C.1}$$

where $\mathbf{x} = (x_{smr} \,\forall\, s, m, r)$ are the model runs and $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_M, \gamma_1, \ldots, \gamma_M)$ is the vector of random effects, and $\boldsymbol{\phi} = \left(\mu, \beta, \lambda, \sigma_H^2, \sigma_F^2, \sigma_\alpha^2, \sigma_\gamma^2\right)$ is the vector of parameters. The prior probability of the random effects, conditional on the parameters is

$$\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\phi}\right) \tau_\alpha^{M/2} \exp\left(-\frac{\tau_\alpha}{2} \sum_{m=1}^{M} \left(\alpha_m - 0\right)^2\right) \tau_{\gamma|\alpha}^{M/2} \exp\left(-\frac{\tau_{\gamma|\alpha}}{2} \sum_{m=1}^{M} \left(\gamma_m - \lambda\alpha_m\right)^2\right) \tag{C.2}$$

and the prior probability of the parameters is

$$\begin{aligned}\Pr\left(\boldsymbol{\phi}\right) \propto \exp\left(-\frac{b_\mu}{2}\left(\mu - a_\mu\right)^2\right) \exp\left(-\frac{b_\beta}{2}\left(\beta - a_\beta\right)^2\right) \exp\left(-\frac{b_\lambda}{2}\left(\lambda - a_\lambda\right)^2\right) \\ \tau_H^{c_H - 1} \exp\left(-d_H \tau_H\right) \tau_F^{c_F - 1} \exp\left(-d_F \tau_F\right) \tau_\alpha^{c_\alpha - 1} \exp\left(-d_\alpha \tau_\alpha\right) \tau_{\gamma|\alpha}^{c_{\gamma|\alpha} - 1} \exp\left(-d_{\gamma|\alpha} \tau_{\gamma|\alpha}\right)\end{aligned} \tag{C.3}$$

so that the joint posterior of the parameters is given by

$$\Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right) \propto \Pr\left(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi}\right)$$

up to a constant of integration. The full conditionals distributions are found by the same method outlined in Appendix B, so that

$$\begin{aligned}\mu|\ldots \sim N\Bigg(&\frac{b_\mu a_\mu + \tau_H \sum_{m=1}^{M} N_{Hm}\left(\overline{x}_{Hm.} - \alpha_m\right) + \tau_F \sum_{m=1}^{M} N_{Fm}\left(\overline{x}_{Fm.} - \alpha_m - \beta - \gamma_m\right)}{b_\mu + N_{H.}\tau_H + N_{F.}\tau_F}, \\ &\left(b_\mu + N_{H.}\tau_H + N_{F.}\tau_F\right)^{-1}\Bigg)\end{aligned} \tag{C.4}$$

$$\beta|\ldots \sim N\left(\frac{b_\beta a_\beta + \tau_F \sum_{m=1}^{M} N_{Fm}\left(\overline{x}_{Fm.} - \mu - \alpha_m - \gamma_m\right)}{b_\beta + N_{F.}\tau_F},\ \left(b_\beta + N_{F.}\tau_F\right)^{-1}\right) \tag{C.5}$$

$$\lambda|\ldots \sim N\left(\frac{b_\lambda a_\lambda + \tau_{\gamma|\alpha} \sum_{m=1}^{M} \alpha_m \gamma_m}{b_\lambda + \tau_{\gamma|\alpha} \sum_{m=1}^{M} \alpha_m^2},\ \left(b_\lambda + \tau_{\gamma|\alpha} \sum_{m=1}^{M} \alpha_m^2\right)^{-1}\right) \tag{C.6}$$

$$\begin{aligned}\alpha_m|\ldots \sim N\Bigg(&\frac{\tau_H N_{Hm}\left(\overline{x}_{Hm.} - \mu\right) + \tau_F N_{Fm}\left(\overline{x}_{Fm.} - \mu - \beta - \gamma_m\right) + \tau_{\gamma|\alpha}\lambda\gamma_m}{N_{Hm}\tau_H + N_{Fm}\tau_F + \tau_\alpha + \lambda^2\tau_{\gamma|\alpha}}, \\ &\left(N_{Hm}\tau_H + N_{Fm}\tau_F + \tau_\alpha + \lambda^2\tau_{\gamma|\alpha}\right)^{-1}\Bigg)\end{aligned} \tag{C.7}$$

$$\gamma_m|\ldots \sim N\left(\frac{\tau_F N_{Fm}\left(\overline{x}_{Fm.}-\mu-\alpha_m-\beta\right)+\tau_{\gamma|\alpha}\lambda\alpha_m}{N_{Fm}\tau_F+\tau_{\gamma|\alpha}}\ ,\ \left(N_{Fm}\tau_F+\tau_{\gamma|\alpha}\right)^{-1}\right) \quad \text{(C.8)}$$

$$\tau_H|\ldots \sim Gamma\left(c_H+\frac{N_{H.}}{2}\ ,\ d_H+\frac{\sum_{m=1}^{M}\sum_{r=1}^{N_{Hm}}\left(x_{Hmr}-\mu-\alpha_m\right)^2}{2}\right) \quad \text{(C.9)}$$

$$\tau_F|\ldots \sim Gamma\left(c_F+\frac{N_{F.}}{2}\ ,\ d_F+\frac{\sum_{m=1}^{M}\sum_{r=1}^{N_{Fm}}\left(x_{Fmr}-\mu-\alpha_m-\beta-\gamma_m\right)^2}{2}\right) \quad \text{(C.10)}$$

$$\tau_\alpha|\ldots \sim Gamma\left(c_\alpha+\frac{M}{2}\ ,\ d_\alpha+\frac{\sum_{m=1}^{M}\left(\alpha_m-0\right)^2}{2}\right) \quad \text{(C.11)}$$

$$\tau_{\gamma|\alpha}|\ldots \sim Gamma\left(c_{\gamma|\alpha}+\frac{M}{2}\ ,\ d_{\gamma|\alpha}+\frac{\sum_{m=1}^{M}\left(\gamma_m-\lambda\alpha_m\right)^2}{2}\right) \quad \text{(C.12)}$$

C.2. Derivation of the full conditional distributions for cross-validation

If the joint posterior distribution of the parameters $\Pr\left(\boldsymbol{\theta},\boldsymbol{\phi}\mid\mathbf{x}\right)$ had a closed form, then it would be simple to add the conditional cross validation procedure in to the marginal procedure in Chapter 4

1. Compute the posterior distribution of the parameters having excluded *all* the runs from model j, i.e., $\Pr\left(\boldsymbol{\theta},\boldsymbol{\phi}\mid\mathbf{x}_{m\neq j}\right)$;

2. Compute the required marginal posterior predictive probabilities from Chapter 4;

3. Update the posterior of the parameters using only the historical runs from model j

$$\Pr\left(\boldsymbol{\theta},\boldsymbol{\phi}\mid\mathbf{x}_{m\neq j},\mathbf{x}_{Hj}\right)\propto\Pr\left(\mathbf{x}_{Hj}\mid\boldsymbol{\theta},\boldsymbol{\phi}\right)\Pr\left(\boldsymbol{\theta},\boldsymbol{\phi}\mid\mathbf{x}_{m\neq j}\right)$$

4. Compute the conditional posterior predictive probability $\Pr\left(\widetilde{\overline{x}_{Fj.}}>\overline{x}_{Fj.}\mid\mathbf{x}_{m\neq j},\mathbf{x}_{Hj}\right)$ defined in Section 5.4.3.

However, we only have N samples from the joint posterior $\Pr(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}_{m\neq j})$ and knowledge of the full conditionals. In order to update for $\mathbf{x}_{Hj}$, a new set of full conditionals must be derived and N new samples obtained from the updated posterior, including thinning and a second burn-in period to allow the chains to converge to their new stationary distributions.

The full conditional distributions for sampling from

$$\Pr(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}_{m\neq j}, \mathbf{x}_{Hj}) \propto \Pr(\mathbf{x}_{m\neq j}, \mathbf{x}_{Hj} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \Pr(\boldsymbol{\theta} \mid \boldsymbol{\phi}) \Pr(\boldsymbol{\phi})$$

are derived by the same method as in Appendix C. Without loss of generality, assume that $j = M$ so that there are runs from M models in the historical scenario and $M - 1$ models in the future scenario, then

$$\mu|\ldots \sim N\left(\frac{b_\mu a_\mu + \tau_H \sum_{m=1}^{M} N_{Hm}(\overline{x}_{Hm.} - \alpha_m) + \tau_F \sum_{m=1}^{M-1} N_{Fm}(\overline{x}_{Fm.} - \alpha_m - \beta - \gamma_m)}{b_\mu + N_{H.}\tau_H + N_{F.}\tau_F}, \right.$$
$$\left. (b_\mu + N_{H.}\tau_H + N_{F.}\tau_F)^{-1}\right) \tag{C.13}$$

where $N_{H.} = \sum_{m=1}^{M} N_{Hm}$ and $N_{F.} = \sum_{m=1}^{M-1} N_{Fm}$

$$\beta|\ldots \sim N\left(\frac{b_\beta a_\beta + \tau_F \sum_{m=1}^{M-1} N_{Fm}(\overline{x}_{Fm.} - \mu - \alpha_m - \gamma_m)}{b_\beta + N_{F.}\tau_F}, \ (b_\beta + N_{F.}\tau_F)^{-1}\right) \tag{C.14}$$

$$\lambda|\ldots \sim N\left(\frac{b_\lambda a_\lambda + \tau_{\gamma|\alpha} \sum_{m=1}^{M-1} \alpha_m \gamma_m}{b_\lambda + \tau_{\gamma|\alpha} \sum_{m=1}^{M-1} \alpha_m^2}, \ \left(b_\lambda + \tau_{\gamma|\alpha} \sum_{m=1}^{M-1} \alpha_m^2\right)^{-1}\right) \tag{C.15}$$

$$\alpha_m|\ldots \sim N\left(\frac{\tau_H N_{Hm}(\overline{x}_{Hm.} - \mu) + \tau_F N_{Fm}(\overline{x}_{Fm.} - \mu - \beta - \gamma_m) + \tau_{\gamma|\alpha}\lambda\gamma_m}{N_{Hm}\tau_H + N_{Fm}\tau_F + \tau_\alpha + \lambda^2\tau_{\gamma|\alpha}}, \right.$$
$$\left. \left(N_{Hm}\tau_H + N_{Fm}\tau_F + \tau_\alpha + \lambda^2\tau_{\gamma|\alpha}\right)^{-1}\right) \quad \text{for} m = 1, \ldots, M-1 \tag{C.16}$$

$$\alpha_M|\ldots \sim N\left(\frac{\tau_H N_{Hm}(\overline{x}_{Hm.} - \mu)}{N_{Hm}\tau_H + \tau_\alpha}, \ (N_{Hm}\tau_H + \tau_\alpha)^{-1}\right) \tag{C.17}$$

$$\gamma_m|\ldots \sim N\left(\frac{\tau_F N_{Fm}\left(\overline{x}_{Fm.}-\mu-\alpha_m-\beta\right)+\tau_{\gamma|\alpha}\lambda\alpha_m}{N_{Fm}\tau_F+\tau_{\gamma|\alpha}}\ ,\ \left(N_{Fm}\tau_F+\tau_{\gamma|\alpha}\right)^{-1}\right)$$
$$\text{for } m=1,\ldots,M-1 \tag{C.18}$$

$$\tau_H|\ldots \sim Gamma\left(c_H+\frac{N_{H.}}{2}\ ,\ d_H+\frac{\sum_{m=1}^{M}\sum_{r=1}^{N_{Hm}}\left(x_{Hmr}-\mu-\alpha_m\right)^2}{2}\right) \tag{C.19}$$

$$\tau_F|\ldots \sim Gamma\left(c_F+\frac{N_{F.}}{2}\ ,\ d_F+\frac{\sum_{m=1}^{M-1}\sum_{r=1}^{N_{Fm}}\left(x_{Fmr}-\mu-\alpha_m-\beta-\gamma_m\right)^2}{2}\right) \tag{C.20}$$

$$\tau_\alpha|\ldots \sim Gamma\left(c_\alpha+\frac{M}{2}\ ,\ d_\alpha+\frac{\sum_{m=1}^{M}\left(\alpha_m-0\right)^2}{2}\right) \tag{C.21}$$

$$\tau_{\gamma|\alpha}|\ldots \sim Gamma\left(c_{\gamma|\alpha}+\frac{M-1}{2}\ ,\ d_{\gamma|\alpha}+\frac{\sum_{m=1}^{M-1}\left(\gamma_m-\lambda\alpha_m\right)^2}{2}\right) \tag{C.22}$$

D. Background to the full framework

D.1. The posterior distribution of the actual historical climate

In Equation 6.19 of Section 6.6, it was shown that the joint posterior of the historical climate of the Earth system and the parameters relating to the ensemble can be factorised as

$$\Pr\left(y_{Ha}, y_H, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}, z\right) = \Pr\left(y_{Ha}, y_H \mid \mu, z\right) \Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right)$$

where $\mathbf{x}$ are the model outputs, z is the observations, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_M, \gamma_1, \ldots \gamma_M)$ is the vector of random effects, and $\boldsymbol{\phi} = (\mu, \beta, \lambda)$ is the vector of parameters.

The distribution of interest is the marginal posterior distribution of the actual historical climate y_H

$$\Pr\left(y_H \mid \mathbf{x}, z\right) = \int\int \Pr\left(y_{Ha}, y_H, \mu \mid \mathbf{x}, z\right) dy_{Ha} d\mu$$

the random effects $\boldsymbol{\theta}$ and the remaining parameters are neglected for brevity. By the application of Bayes' rule and the law of conditional probability, the joint posterior can be decomposed as follows

$$\begin{aligned}
\Pr\left(y_{Ha}, y_H, \mu \mid \mathbf{x}, z\right) &= \Pr\left(y_{Ha}, y_H \mid \mu, z\right) \Pr\left(\mu \mid \mathbf{x}\right) \\
&\propto \Pr\left(\mu, z \mid y_{Ha}, y_H\right) \Pr\left(y_{Ha}, y_H\right) \Pr\left(\mu \mid \mathbf{x}\right) \\
&\propto \Pr\left(z \mid y_{Ha}, y_H, \mu\right) \Pr\left(\mu \mid y_{Ha}, y_H\right) \Pr\left(y_{Ha}, y_H\right) \Pr\left(\mu \mid \mathbf{x}\right) \\
&\propto \Pr\left(z \mid y_{Ha}\right) \Pr\left(\mu \mid y_{Ha}, y_H\right) \Pr\left(y_{Ha}, y_H\right) \Pr\left(\mu \mid \mathbf{x}\right)
\end{aligned}$$

since the observations z are independent of y_H and μ given knowledge of the historical climate that we experienced y_{Ha} (Figure 6.1). Then applying Bayes' rule and

factorising once more yields

$$\begin{aligned}\Pr(y_{Ha}, y_H, \mu \mid \mathbf{x}, z) &\propto \Pr(z \mid y_{Ha}) \Pr(y_{Ha}, y_H \mid \mu) \Pr(\mu \mid \mathbf{x}) \\ &\propto \Pr(z \mid y_{Ha}) \Pr(y_{Ha} \mid y_H, \mu) \Pr(y_H \mid \mu) \Pr(\mu \mid \mathbf{x}) \\ &\propto \Pr(z \mid y_{Ha}) \Pr(y_{Ha} \mid y_H) \Pr(y_H \mid \mu) \Pr(\mu \mid \mathbf{x}) \qquad \text{(D.1)}\end{aligned}$$

since y_{Ha} is independent of μ given y_H (Figure 6.1).

Integrating Equation D.1 over y_{Ha} and μ yields

$$\Pr(y_H \mid \mathbf{x}, z) \propto \Pr(z \mid y_H) \Pr(y_H \mid \mathbf{x}) \qquad \text{(D.2)}$$

The first term, $\Pr(z \mid y_H)$, is the likelihood of the observations given the actual historical climate. The second term, $\Pr(y_H \mid \mathbf{x})$, is the posterior distribution of the actual historical climate given only the model outputs. These are easily obtained from Equation D.1 by noting that $\Pr(z \mid y_{Ha})$ is simply Equation 6.6, and $\Pr(y_{Ha} \mid y_H)$ is given by Equation 6.4a. Since both components are normally distributed, we can immediately write

$$\Pr(z \mid y_H) \sim N\left(y_H \ , \ \sigma_{Ha}^2 + \sigma_z^2\right) \qquad \text{(D.3)}$$

Similarly, $\Pr(y_H \mid \mu)$ in Equation D.1 is simply Equation 6.1b. The second term, $\Pr(\mu \mid \mathbf{x})$, is the marginal posterior distribution of the expected climate of the ensemble μ. We assume that $\Pr(\mu \mid \mathbf{x})$ is known (e.g., from Equation C.4 in Appendix C.1), and well approximated by

$$\mu \mid \mathbf{x} \sim N\left(\mu_\mu, \sigma_\mu^2\right)$$

Since both $\Pr(y_H \mid \mu)$ and $\Pr(\mu \mid \mathbf{x})$ have normal densities, we can immediately write

$$\Pr(y_H \mid \mathbf{x}) \sim N\left(\mu_\mu \ , \ \sigma_\mu^2 + \sigma_{\Delta_H}^2\right) \qquad \text{(D.4)}$$

The posterior distribution of the actual historical climate given both the model outputs and the observations is then easily obtained from Equations D.2, D.3 and D.4

$$\Pr(y_H \mid \mathbf{x}, z) \sim N\left(\frac{\tau_{y|x}\mu_\mu + \tau_{z|y} z}{\tau_{y|x} + \tau_{z|y}} \ , \ \left(\tau_{y|x} + \tau_{z|y}\right)^{-1}\right)$$

where

$$\tau_{z|y} = \left(\sigma_{Ha}^2 + \sigma_z^2\right)^{-1} \quad \text{and} \quad \tau_{y|x} = \left(\sigma_\mu^2 + \sigma_{\Delta_H}^2\right)^{-1}$$

D.2. Obtaining identical inferences from different assumptions

It can be shown that the frameworks proposed by Rougier et al. (2013) and Chandler (2013), reviewed in Chapter 2, will produce identical inferences about the actual climate, provided that identical distributional assumptions are made for key components. The goal is to make inferences about the climate of the Earth system Y, given the data from the models $\mathbf{X}$, and the observations Z. The distribution of interest is

$$\Pr\left(Y \mid \mathbf{X}, Z\right) = \int \Pr\left(Y, \mathrm{M}\left(X\right) \mid \mathbf{X}, Z\right) d\,\mathrm{M}\left(X\right) \tag{D.5}$$

The R_m, W, U and B terms are all treated as zero mean random departures, and so are neglected for brevity. In the framework proposed by Rougier et al. (2013), the joint posterior of the actual climate Y and the model consensus $\mathrm{M}\left(X\right)$ can be decomposed as follows

$$\begin{aligned}
\Pr\left(Y, \mathrm{M}\left(X\right) \mid \mathbf{X}, Z\right) &\propto \Pr\left(\mathbf{X}, Z \mid Y, \mathrm{M}\left(X\right)\right) \Pr\left(Y, \mathrm{M}\left(X\right)\right) \\
&\propto \Pr\left(Z \mid \mathbf{X}, Y, \mathrm{M}\left(X\right)\right) \Pr\left(\mathbf{X} \mid Y, \mathrm{M}\left(X\right)\right) \Pr\left(Y, \mathrm{M}\left(X\right)\right) \\
&\propto \Pr\left(Z \mid Y\right) \Pr\left(\mathbf{X} \mid \mathrm{M}\left(X\right)\right) \Pr\left(Y, \mathrm{M}\left(X\right)\right) \\
&\propto \Pr\left(Z \mid Y\right) \Pr\left(\mathbf{X} \mid \mathrm{M}\left(X\right)\right) \Pr\left(Y \mid \mathrm{M}\left(X\right)\right) \Pr\left(\mathrm{M}\left(X\right)\right)
\end{aligned} \tag{D.6}$$

The first step is a simple application of Bayes' theorem, and the second line follows from the law of conditional probability. The simplifications in the third line are most easily understood from Figure 6.3. From the graph, it is clear that Z is independent of $\mathbf{X}$ and $\mathrm{M}\left(X\right)$ given Y, and similarly $\mathbf{X}$ is independent of Y given $\mathrm{M}\left(X\right)$. The final line follows from the law of conditional probability. The decomposition for the generalised "truth plus error" framework of Chandler (2013) differs only in the final line where

$$\Pr\left(Y, \mathrm{M}\left(X\right) \mid \mathbf{X}, Z\right) \propto \Pr\left(Z \mid Y\right) \Pr\left(\mathbf{X} \mid \mathrm{M}\left(X\right)\right) \Pr\left(\mathrm{M}\left(X\right) \mid Y\right) \Pr\left(Y\right) \tag{D.7}$$

which follows by applying the law of conditional probability in the opposite direction.

The two factorisations share two common terms. The $\Pr\left(Z \mid Y\right)$ term is simply the likelihood of the observations, given the actual climate (Equations 2.14a and 2.16a). The $\Pr\left(\mathbf{X} \mid \mathrm{M}\left(X\right)\right)$ term is the likelihood of the model outputs, conditional on the model consensus (Equations 2.14c and 2.16c). It is reasonable to suppose that we would make the same judgements about both likelihoods regardless of which framework we adopt. The remaining terms in the posterior decompositions differ between the two frameworks. The $\Pr\left(Y \mid \mathrm{M}\left(X\right)\right)$ and $\Pr\left(\mathrm{M}\left(X\right) \mid Y\right)$ terms specify our judgements about the discrepancies U and B, respectively. The $\Pr\left(\mathrm{M}\left(X\right)\right)$ and

$\Pr(Y)$ terms are the prior probabilities assigned to the model consensus and the actual climate. Both Rougier et al. (2013) and Chandler (2013) specify symmetric discrepancies and vague priors for $\mathrm{M}(X)$ and Y as the default choices. If identical symmetric distributions are specified for $\Pr(Y \mid \mathrm{M}(X))$ and $\Pr(\mathrm{M}(X) \mid Y)$, and uniform priors are assumed for $\mathrm{M}(X)$ and Y, then the inferences obtained from the two frameworks will be identical.

D.3. The alternative parameterisation of Tebaldi et al. (2005)

The alternative parameterisation of Tebaldi et al. (2005) including the extension by Smith et al. (2009) is

$$\begin{aligned} x_{Hm} &\sim N\left(y_H \ , \ \tau_m^{-1}\right) \\ x_{Fm} \mid x_{Hm} &\sim N\left(x_{Hm} + y_R + \lambda\left(x_{Hm} - y_H\right) \ , \ \left(\theta\tau_m\right)^{-1}\right) \\ \tau_m &\overset{iid}{\sim} Gamma\left(k \ , \ l\right) \\ z \mid y_H &\sim N\left(y_H \ , \ \tau_z^{-1}\right) \end{aligned}$$

Normal priors are assumed for y_H, y_R and λ, and Gamma priors for θ, k and l. The joint posterior distribution of the parameters is given by

$$\begin{aligned} \Pr\left(y_H, y_R, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{x}\right) \propto\ & \tau_z^{1/2} \exp\left(-\frac{1}{2}\tau_z\left(z - y_H\right)^2\right) \\ & \left(\prod_{m=1}^{M} \tau_m^{1/2}\right) \exp\left(-\frac{1}{2}\sum_{m=1}^{M} \tau_m\left(x_{Hm} - y_H\right)^2\right) \\ & \theta^{M/2}\left(\prod_{m=1}^{M} \tau_m^{1/2}\right) \exp\left(-\frac{1}{2}\sum_{m=1}^{M} \theta\tau_m\left(x_{Fm} - x_{Hm} - y_R - \lambda\left(x_{Hm} - y_H\right)\right)^2\right) \\ & \left(\frac{l^k}{\Gamma(k)}\right)^M \left(\prod_{m=1}^{M} \tau_m\right)^{k-1} \exp\left(-l\sum_{m=1}^{M} \tau_m\right) \\ & \exp\left(-\frac{b_{y_H}}{2}\left(y_H - a_{y_H}\right)^2\right) \exp\left(-\frac{b_{y_R}}{2}\left(y_R - a_{y_R}\right)^2\right) \exp\left(-\frac{b_\lambda}{2}\left(\lambda - a_\lambda\right)^2\right) \\ & \theta^{c_\theta - 1} \exp\left(-d_\theta\theta\right) k^{c_k - 1} \exp\left(-d_k k\right) l^{c_l - 1} \exp\left(-d_l l\right) \end{aligned}$$

up to a constant of integration, where $\mathbf{x} = (x_{smr} \ \forall \ s, m, r)$ are the model runs and $\boldsymbol{\theta} = (\tau_1, \ldots, \tau_M)$ is the vector of model specific precisions, and $\boldsymbol{\phi} = (\lambda, \theta, k, l)$ is the vector of parameters. The full conditional distributions are found by the same

method outlined in Appendix B, so that

$$y_H \mid \ldots \sim N\left(\frac{b_{y_H}a_{y_H} + \tau_z z + \sum_{m=1}^{M}\tau_m x_{Hm} - \lambda\theta\sum_{m=1}^{M}\tau_m\left(x_{Fm} - x_{Hm} - y_R - \lambda x_{Hm}\right)}{b_{y_H} + \tau_z + \sum_{m=1}^{M}\tau_m + \theta\lambda^2\sum_{m=1}^{M}\tau_m}\ ,\right.$$
$$\left.\left(b_{y_H} + \tau_z + \sum_{m=1}^{M}\tau_m + \theta\lambda^2\sum_{m=1}^{M}\tau_m\right)^{-1}\right) \tag{D.8}$$

$$y_R \mid \ldots \sim N\left(\frac{b_{y_R}a_{y_R} + \theta\sum_{m=1}^{M}\tau_m\left(x_{Fm} - x_{Hm} - \lambda\left(x_{Hm} - y_H\right)\right)}{b_{y_R} + \theta\sum_{m=1}^{M}\tau_m}\ ,\ \left(b_{y_R} + \theta\sum_{m=1}^{M}\tau_m\right)^{-1}\right) \tag{D.9}$$

$$\lambda \mid \ldots \sim N\left(\frac{b_\lambda a_\lambda + \theta\sum_{m=1}^{M}\tau_m\left(x_{Hm} - y_H\right)\left(x_{Fm} - x_{Hm} - y_R\right)}{b_\lambda + \theta\sum_{m=1}^{M}\tau_m\left(x_{Hm} - y_H\right)^2}\ ,\right.$$
$$\left.\left(b_\lambda + \theta\sum_{m=1}^{M}\tau_m\left(x_{Hm} - y_H\right)^2\right)^{-1}\right) \tag{D.10}$$

$$\theta \mid \ldots \sim Gamma\left(c_\theta + \frac{M}{2}\ ,\ d_\theta + \frac{\theta\sum_{m=1}^{M}\left(x_{Fm} - x_{Hm} - y_R - \lambda\left(x_{Hm} - y_H\right)\right)^2}{2}\right) \tag{D.11}$$

$$\tau_m \mid \ldots \sim Gamma\left(k+1\ ,\ l + \frac{\left(x_{Hm} - y_H\right)^2}{2} + \frac{\theta\left(x_{Fm} - x_{Hm} - y_R - \lambda\left(x_{Hm} - y_H\right)\right)^2}{2}\right) \tag{D.12}$$

The conditional distributions of k and l do not take the form of any standard probability distributions. Instead, a Metropolis step can be used to update those parameters, as described by Smith et al. (2009).

Bibliography

Abe, M., H. Shiogama, T. Nozawa, and S. Emori, 2011: Estimation of future surface temperature changes constrained using the future-present correlated modes in inter-model variability of CMIP3 multimodel simulations. *Journal of Geophysical Research: Atmospheres*, **116**, D18 104, doi:10.1029/2010JD015111.

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723, doi:10.1109/TAC.1974.1100705.

Allen, M. R. and W. J. Ingram, 2002: Constraints on future changes in climate and the hydrologic cycle. *Nature*, **419**, 224–232, doi:10.1038/nature01092.

Allen, M. R., J. F. B. Mitchell, and P. A. Stott, 2013: Test of a decadal climate forecast. *Nature Geoscience*, **6 (4)**, 243–244, doi:10.1038/ngeo1788.

Allen, M. R., P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, **407**, 617–620, doi:10.1038/35036559.

Alley, R. B. and Coauthors, 2007: Summary for Policymakers. *Climate Change 2007: The Physical Science Basis*, S. Solomon and et al., Eds., Cambridge University Press, 1–18.

Anderson, D., K. I. Hodges, and B. J. Hoskins, 2003: Sensitivity of feature-based analysis methods of storm tracks to the form of background field removal. *Monthly Weather Review*, **131 (3)**, 565–573, doi:10.1175/1520-0493(2003)131⟨0565:SOFBAM⟩2.0.CO;2.

Annan, J. D. and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophysical Research Letters*, **37**, L02 703, doi:10.1029/2009GL041994.

Annan, J. D. and J. C. Hargreaves, 2011: Understanding the CMIP3 multimodel ensemble. *Journal of Climate*, **24**, 4529–4538, doi:10.1175/2011JCLI3873.1.

Barnett, T. H., et al., 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *Journal of Climate*, **18 (9)**, 1291–1314, doi:10.1175/JCLI3329.1.

Bengtsson, L., K. I. Hodges, and N. Keenlyside, 2009: Will extratropical storms intensify in a warmer climate? *Journal of Climate*, **22 (9)**, 2276–2301, doi:10.1175/2008JCLI2678.1.

Bengtsson, L., K. I. Hodges, and E. Roeckner, 2006: Storm tracks and climate change. *Journal of Climate*, **19 (15)**, 3518–3543, doi:10.1175/JCLI3815.1.

Berger, J. O., 1985: *Statistical decision theory and Bayesian analysis*. 2d ed., Springer, 618 pp.

Berliner, L. M. and Y. Kim, 2008: Bayesian design and analysis for superensemble-based climate forecasting. *Journal of Climate*, **21**, 1891–1910, doi:10.1175/2007JCLI1619.1.

Bernardo, J. M. and A. F. M. Smith, 2000: *Bayesian Theory*. Wiley, 610 pp.

Besag, J., 1974: Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **36**, 192–236, URL `http://www.jstor.org/stable/2984812`.

Bhend, J. and P. Whetton, 2013: Effective constraints for regional climate change projections. *12th International Meeting on Statistical Climatology*.

Bishop, C. H. and G. Abramowitz, 2013: Climate model dependence and the replicate Earth paradigm. *Climate Dynamics*, **41**, 885–900, doi:10.1007/s00382-012-1610-y.

Blackmon, M. L., 1976: A climatological spectral study of the 500mb geopotential height of the Northern Hemisphere. *Journal of the Atmospheric Sciences*, **33 (8)**, 1607–1623, doi:10.1175/1520-0469(1976)033⟨1607:ACSSOT⟩2.0.CO;2.

Blender, R., K. Fraedrich, and F. Lunkeit, 1997: Identification of cyclone-track regimes in the North Atlantic. *Quarterly Journal of the Royal Meteorological Society*, **123**, 727–741, doi:10.1002/qj.49712353910.

Boé, J., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nature Geoscience*, **2 (5)**, 341–343, doi:10.1038/ngeo467.

Bony, S., et al., 2006: How well do we understand and evaluate climate change feedback processes? *Journal of Climate*, **19 (15)**, 3445–3482, doi:10.1175/JCLI3819.1.

Boucher, O., et al., 2013: Clouds and aerosols. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker, D. Qin, G.-K. Plattner, M. M. B. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Eds., Cambridge University Press.

Bracegirdle, T. J. and D. B. Stephenson, 2012: Higher precision estimates of regional polar warming by ensemble regression of climate model projections. *Climate Dynamics*, **39 (12)**, 2805–2821, doi:10.1007/s00382-012-1330-3.

Bracegirdle, T. J. and D. B. Stephenson, 2013: On the robustness of emergent constraints used in multimodel climate change projections of Arctic warming. *Journal of Climate*, **26 (2)**, 669–678, doi:10.1175/JCLI-D-12-00537.1.

Brown, M. B. and A. B. Forsythe, 1974: Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**, 364–367, URL `http://www.jstor.org/stable/2285659`.

Buser, C. M., H. R. Künsch, D. Lüthi, M. Wild, and C. Schär, 2009: Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Climate Dynamics*, **33 (6)**, 849–868, doi:10.1007/s00382-009-0588-6.

Buser, C. M., H. R. Künsch, and A. Weber, 2010: Biases and uncertainty in climate projections. *Scandinavian Journal of Statistics*, **37 (2)**, 179–199, doi:10.1111/j.1467-9469.2009.00686.x.

Catto, J. L., L. C. Shaffrey, and K. I. Hodges, 2010: Can climate models capture the structure of extratropical cyclones? *Journal of Climate*, **23**, 1621–1635, doi:10.1175/2009JCLI3318.1.

Chandler, R., 2003: RANDGEN. URL `http://www.homepages.ucl.ac.uk/~ucakarc/work/randgen.html`.

Chandler, R. E., 2013: Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A*, **371**, doi:10.1098/rsta.2012.0388.

Chang, E. K. M., Y. Guo, and X. Xia, 2012: CMIP5 multimodel ensemble projection of storm track change under global warming. *Journal of Geophysical Research: Atmospheres*, **117**, D23 118, doi:10.1029/2012JD018578.

Chang, E. K. M., Y. Guo, X. Xia, and M. Zheng, 2013: Storm-track activity in IPCC AR4/CMIP3 model simulations. *Journal of Climate*, **26**, 246–260, doi:10.1175/JCLI-D-11-00707.1.

Charney, J. G., 1947: The dynamics of long waves in a baroclinic westerly current. *Journal of Meteorology*, **4 (5)**, 136–162, doi:10.1175/1520-0469(1947)004⟨0136:TDOLWI⟩2.0.CO;2.

Christensen, J. H., F. Boberg, O. B. Christensen, and P. Lucas-Picher, 2008: On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, **35**, L20 709, doi:10.1029/2008GL035694.

Christensen, J. H., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen, 2010: Weight assignment in regional climate models. *Climate Research*, **44**, 179–194, doi:10.3354/cr00916.

Christensen, J. H., et al., 2013: Climate phenomena and their relevance for future regional climate change. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker, D. Qin, G.-K. Plattner, M. M. B. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Eds., Cambridge University Press.

Cohen, J., 1973: Eta-squared and partial Eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, **33**, 107–112, doi:10.1177/001316447303300111.

Cohen, J., 1988: *Statistical Power Analysis for the Behavioral Sciences*. 2d ed., Routledge, 590 pp.

Colle, B. A., Z. Zhang, K. A. Lombardo, E. Chang, P. Liu, and M. Zhang, 2013: Historical evaluation and future prediction of Eastern North American and Western Atlantic extratropical cyclones in the CMIP5 models during the cool season. *Journal of Climate*, **26**, 6882–6903, doi:10.1175/JCLI-D-12-00498.1.

Collins, M., B. Booth, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2006a: Towards quantifying uncertainty in transient climate change. *Climate Dynamics*, **27 (2-3)**, 127–147, doi:10.1007/s00382-006-0121-0.

Collins, M., R. E. Chandler, P. M. Cox, J. M. Huthnance, J. Rougier, and D. B. Stephenson, 2012: Quantifying future climate change. *Nature Climate Change*, **2 (6)**, 403–409, doi:10.1038/nclimate1414.

Collins, M., et al., 2006b: Interannual to decadal climate predictability in the North Atlantic: A multimodel-ensemble study. *Journal of Climate*, **19**, 1195–1203, doi:10.1175/JCLI3654.1.

Collins, M., et al., 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker, D. Qin, G.-K. Plattner, M. M. B. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Eds., Cambridge University Press.

Compo, G. P., et al., 2011: The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, **137 (654)**, 1–28, doi:10.1002/qj.776.

Connolley, W. M. and T. J. Bracegirdle, 2007: An Antarctic assessment of IPCC AR4 coupled models. *Geophysical Research Letters*, **34**, L22 505, doi:10.1029/2007GL031648.

Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jone, and C. M. Luke, 2013: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature*, **494**, 341–344, doi:10.1038/nature11882.

Craig, P. S., M. Goldstein, J. C. Rougier, and A. H. Seheult, 2001: Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, **96**, 717–729, URL `http://www.jstor.org/stable/2670309`.

Cressie, N. A. C., 1993: *Statistics for Spatial Data*. Revised edition ed., Wiley, 928 pp.

Cubasch, U., G. Meehl, G. Boer, M. Stouffer, R.J. Dix, A. Noda, C. Senior, S. Raper, and K. Yap, 2001: Projections of future climate change. *Climate Change 2001: The Scientific Basis*, J. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., Cambridge University Press.

Daron, J. D. and D. A. Stainforth, 2013: On predicting climate under climate change. *Environmental Research Letters*, **8**, 034 021, doi:10.1088/1748-9326/8/3/034021.

Davison, A. C., 2003: *Statistical Models*. Cambridge University Press, 726 pp.

Dee, D., J. Fasullo, D. Shea, and J. Walsh, 2014: The climate data guide: Atmospheric reanalysis: Overview & comparison tables. URL `https://climatedataguide.ucar.edu/climate-data/atmospheric-reanalysis-overview-comparison-tables`, accessed on 4 Nov 2013.

Dee, D. P., et al., 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597, doi:10.1002/qj.828.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *Journal of Climate*, **20 (12)**, 2810–2826, doi:10.1175/JCLI4179.1.

Deser, C., R. Knutti, S. Solomin, and A. S. Phillips, 2012a: Communication of the role of natural variability in future North American climate. *Nature Climate Change*, **2**, 775779, doi:10.1038/NCLIMATE1562.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012b: Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, **38**, 527–546, doi:10.1007/s00382-010-0977-x.

Doblas-Reyes, F. J., V. Pavan, and D. B. Stephenson, 2003: The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate Dynamics*, **21 (5-6)**, 501–514, doi:10.1007/s00382-003-0350-4.

Doksum, K. A. and G. L. Sievers, 1976: Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, **63 (3)**, 421–434, URL `http://www.jstor.org/stable/2335720`.

Eady, E. T., 1949: Long waves and cyclone waves. *Tellus*, **1 (3)**, 33–52, doi:10.1111/j.2153-3490.1949.tb01265.x.

Eyring, V., et al., 2007: Multimodel projections of stratospheric ozone in the 21st century. *Journal of Geophysical Research: Atmospheres*, **112**, D16 303, doi:10.1029/2006JD008332.

Fasullo, J. T. and K. E. Trenberth, 2012: A less cloudy future: The role of subtropical subsidence in climate sensitivity. *Science*, **338**, 792–794, doi:10.1126/science.1227465.

Ferro, C. A. T., 2004: Attributing variation in a regional climate change modelling experiment. Tech. rep., EU Project PRUDENCE. URL `http://prudence.dmi.dk/public/publications/analysis_of_variance.pdf`.

Flato, G., et al., 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker, D. Qin, G.-K. Plattner, M. M. B. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Eds., Cambridge University Press, chap. 9, 741–866.

Forest, C. E., P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster, 2002: Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, **295 (5552)**, 113–117, doi:10.1126/science.1064419, `http://www.sciencemag.org/content/295/5552/113.full.pdf`.

Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka, and G. A. Meehl, 2007a: Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophysical Research Letters*, **34**, L06 711, doi:10.1029/2006GL027754.

Furrer, R., S. R. Sain, D. Nychka, and G. A. Meehl, 2007b: Multivariate Bayesian analysis of atmosphereocean general circulation models. *Environmental and Ecological Statistics*, **14**, 249–266, doi:10.1007/s10651-007-0018-z.

Fyfe, J. C., N. P. Gillett, and F. W. Zwiers, 2013: Overestimated global warming over the past 20 years. *Nature Climate Change*, **3**, 767–769, doi:10.1038/nclimate1972.

Garthwaite, P., I. Jolliffe, and B. Jones, 2002: *Statistical Inference*. 2d ed., Oxford University Press, 342 pp.

Gastineau, G. and B. J. Soden, 2009: Model projected changes of extreme wind events in response to global warming. *Geophysical Research Letters*, **36**, L10 810, doi:10.1029/2009GL037500.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, 2014: *Bayesian Data Analysis*. 3d ed., Chapman & Hall / CRC, 675 pp.

Geng, Q. and M. Sugi, 2003: Possible change of extratropical cyclone activity due to enhanced greenhouse gases and sulfate aerosols - study with a high-resolution AGCM. *Journal of Climate*, **16 (13)**, 2262–2274, doi:10.1175/1520-0442(2003)16⟨2262:PCOECA⟩2.0.CO;2.

Giorgi, F. and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method. *Journal of Climate*, **15 (10)**, 1141–1158, doi:10.1175/1520-0442(2002)015⟨1141:COAURA⟩2.0.CO;2.

Giorgi, F. and L. O. Mearns, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophysical Research Letters*, **30 (12)**, doi:10.1029/2003GL017130.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, **113**, D06 104, doi:10.1029/2007JD008972.

Goldstein, M. and D. Wooff, 2007: *Bayes Linear Statisics, Theory & Methods*. Wiley, 536 pp.

Greene, A. M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *Journal of Climate*, **19**, 4326–4343, doi:10.1175/JCLI3864.1.

Gregory, J. M., et al., 2005: A model intercomparison of changes in the Atlantic thermohaline circulation in response to increasing atmospheric CO2 concentration. *Geophysical Research Letters*, **32 (12)**, L12 703, doi:10.1029/2005GL023209.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting I. Basic concept. *Tellus A*, **57**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hall, A. and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters*, **33 (3)**, L03 502, doi:10.1029/2005GL025127.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Reviews of Geophysics*, **48**, doi:10.1029/2010RG000345.

Harris, G. R., D. M. H. Sexton, B. B. B. Booth, M. Collins, and J. M. Murphy, 2013: Probabilistic projections of transient climate change. *Climate Dynamics*, **40**, 2937–2972, doi:10.1007/s00382-012-1647-y.

Harvey, B. J., L. C. Shaffrey, and T. J. Woolings, 2013: Equator-to-pole temperature differences and the extra-tropical storm track responses of the CMIP5 climate models. *Climate Dynamics*, doi:10.1007/s00382-013-1883-9.

Hawkins, E. and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, **90**, 1095–1107, doi:10.1175/2009BAMS2607.1.

Hegerl, G. C., et al., 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press.

Hingray, B., A. Mezghani, and T. A. Buishand, 2007: Development of probability distributions for regional climate change from uncertain global mean warming and an uncertain scaling relationship. *Hydrology and Earth System Sciences*, **11 (3)**, 1097–1114, doi:10.5194/hess-11-1097-2007.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Monthly Weather Review*, **122 (11)**, 2573–2585, doi:10.1175/1520-0493(1994)122⟨2573:AGMFTA⟩2.0.CO;2.

Hodges, K. I., 1995: Feature tracking on the unit sphere. *Monthly Weather Review*, **123 (12)**, 3458–3465, doi:10.1175/1520-0493(1995)123⟨3458:FTOTUS⟩2.0.CO;2.

Hodges, K. I., 1996: Spherical nonparametric estimators applied to the UGAMP model integration for AMIP. *Monthly Weather Review*, **124 (12)**, 2914–2932, doi:10.1175/1520-0493(1996)124⟨2914:SNEATT⟩2.0.CO;2.

Hodges, K. I., 1999: Adaptive constraints for feature tracking. *Monthly Weather Review*, **127 (6)**, 1362–1373, doi:10.1175/1520-0493(1999)127⟨1362:ACFFT⟩2.0.CO;2.

Hodges, K. I., R. W. Lee, and L. Bengtsson, 2011: A comparison of extratropical cyclones in recent reanalyses ERA-Interim, NASA MERRA, NCEP CFSR, and JRA-25. *Journal of Climate*, **24 (18)**, 4888–4906, doi:10.1175/2011JCLI4097.1.

Hoerl, A. E. and R. W. Kennard, 1970: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12 (1)**, 55–67, doi:10.1080/00401706.1970.10488634, http://www.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Statistical Science*, **14 (4)**, 382–401, URL http://www.jstor.org/stable/2676803.

Holland, M. M. and C. M. Bitz, 2003: Polar amplification of climate change in coupled models. *Climate Dynamics*, **21**, 221–232, doi:10.1007/s00382-003-0332-6.

Hoskins, B. J. and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *Journal of the Atmospheric Sciences*, **59 (6)**, 1041–1061, doi:10.1175/1520-0469(2002)059⟨1041:NPOTNH⟩2.0.CO;2.

Houghton, J., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, (Eds.) , 2001: *Climate Change 2001: The Scientific Basis*. Cambridge University Press, 881 pp., URL http://www.grida.no/publications/other/ipcc_tar/.

Howson, C. and P. Urbach, 1993: *Scientific reasoning : the Bayesian approach*. 2d ed., Open Court, 470 pp.

Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti, 2011: Constraints on climate sensitivity from radiation patterns in climate models. *Journal of Climate*, **24**, 1034–1052, doi:10.1175/2010JCLI3403.1.

Ingram, W., 2010: A very simple model for the water vapour feedback on climate change. *Quarterly Journal of the Royal Meteorological Society*, **136 (646)**, 30–40, doi:10.1002/qj.546.

Jansen, E., et al., 2007: Palaeoclimate. *Climate Change 2007: The Physical Science Basis*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press.

Johnson, N. L. and S. Kotz, 1970: *Continuous Univariate Distributions*, Vol. 2. 1st ed., John Wiley & Sons, 306 pp.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1995: *Continuous Univariate Distributions*, Vol. 2. 2d ed., John Wiley & Sons, Ltd, 784 pp.

Jun, M., R. Knutti, and D. W. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence. *Journal of the American Statistical Association*, **103**, 934–947, doi:10.1198/016214507000001265.

Kang, E. L. and N. Cressie, 2013: Bayesian hierarchical ANOVA of regional climate-change projections from NARCCAP Phase II. *International Journal of Applied Earth Observation and Geoinformation*, **22**, 3–15, doi:10.1016/j.jag.2011.12.007.

Karpechko, A. Y., D. Maraun, and V. Eyring, 2013: Improving Antarctic total ozone projections by a process-oriented multiple diagnostic ensemble regression. *Journal of the Atmospheric Sciences*, **70**, 3959–3976, doi:10.1175/JAS-D-13-071.1.

Katz, R. W., P. F. Craigmile, P. Guttorp, M. Haran, B. Sansó, and M. L. Stein, 2013: Uncertainty analysis in climate change assessments. *Nature Climate Change*, **3**, 769–771, doi:10.1038/nclimate1980.

Kennedy, M. C. and A. O'Hagan, 2001: Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 425–464, doi:10.1111/1467-9868.00294.

Kettleborough, J. A., B. B. B. Booth, P. A. Stott, and M. R. Allen, 2007: Estimates of uncertainty in predictions of global mean surface temperature. *Journal of Climate*, **20**, 843–855, doi:10.1175/JCLI4012.1.

Kharin, V. V. and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *Journal of Climate*, **15 (7)**, 793–799, doi:10.1175/1520-0442(2002)015⟨0793:CPWME⟩2.0.CO;2.

Knutti, R., 2008a: Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A*, **366 (1885)**, 4674–4664, doi:10.1098/rsta.2008.0169.

Knutti, R., 2008b: Why are climate models reproducing the observed global surface warming so well? *Geophysical Research Letters*, **35 (18)**, L18 704, doi:10.1029/2008GL034932.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns, 2010a: Good practice guidance paper on assessing and combining multi model climate projections. *Meeting report of the Intergovernmental Panel On Climate Change Expert Meeting on Assessing and Combining Multiple Model Climate Projections*, T. Stocker, Q. Dahe, G.-K. Plattner, M. Tignor, and P. Midgley, Eds., IPCC Working Group I Technical Support Unit, URL `http://www.ipcc.ch/pdf/supporting-material/IPCC_EM_MME_GoodPracticeGuidancePaper.pdf`.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010b: Challenges in combining projections from multiple climate models. *Journal of Climate*, **23 (10)**, 2739–2758, doi:10.1175/2009JCLI3361.1.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, **40**, 1194–1199, doi:10.1002/grl.50256.

Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *Journal of Climate*, **19**, 4224–4233, doi:10.1175/JCLI3865.1.

Knutti, R., T. F. Socker, F. Joos, and G.-K. Plattner, 2002: Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature*, **416**, 719–723, doi:10.1038/416719a.

Knutti, R., et al., 2008: A review of uncertainties in global temperature projections over the twenty-first century. *Journal of Climate*, **21 (11)**, 2651–2663, doi:10.1175/2007JCLI2119.1.

König, W., S. R., and F. Sielmann, 1993: Objective identification of cyclones in GCM simulations. *Journal of Climate*, **6 (12)**, 2217–2231, doi:10.1175/1520-0442(1993)006⟨2217:OIOCIG⟩2.0.CO;2.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285 (5433)**, 1548–1550, doi:10.1126/science.285.5433.1548.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, and E. Williford, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, **13 (23)**, 4196–4216, doi:10.1175/1520-0442(2000)013⟨4196:MEFFWA⟩2.0.CO;2.

Krzanowski, W. J., 1998: *An Introduction to Statistical Modelling*. John Wiley & Sons, Ltd, 264 pp.

Lambert, S. J. and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, **17 (2-3)**, 83–106, doi:10.1007/PL00013736.

Lambert, S. J. and J. C. Fyfe, 2006: Changes in winter cyclone frequencies and strengths simulated in enhanced greenhouse warming experiments: results from the models participating in the IPCC diagnostic exercise. *Climate Dynamics*, **26 (7-8)**, 713–728, doi:10.1007/s00382-006-0110-3.

Lambert, S. J., J. Sheng, and J. Boyle, 2002: Winter cyclone frequencies in thirteen models participating in the Atmospheric Model Intercomparison Project (AMIP1). *Climate Dynamics*, **19 (1)**, 1–16, doi:10.1007/s00382-001-0206-8.

Livezey, R. E. and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review*, **111 (1)**, 46–59, doi:10.1175/1520-0493(1983)111%3C0046:SFSAID%3E2.0.CO;2.

Lopez, A., C. Tebaldi, M. New, D. Stainforth, M. Allen, and J. Kettleborough, 2006: Two approaches to quantifying uncertainty in global temperature changes. *Journal of Climate*, **19**, 4785–4796, doi:10.1175/JCLI3895.1.

Mahlstein, I. and R. Knutti, 2011: Ocean heat transport as a cause for model uncertainty in projected Arctic warming. *Journal of Climate*, **24**, 1451–1460, doi:10.1175/2010JCLI3713.1.

Masson, D. and R. Knutti, 2011: Climate model genealogy. *Geophysical Research Letters*, **38**, L08 703, doi:10.1029/2011GL046864.

Mastrandrea, M. D., et al., 2010: Guidance note for lead authors of the IPCC Fifth Assessment Report on consistent treatment of uncertainties. Tech. rep., Intergovernmental Panel on Climate Change (IPCC). URL `http://www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf`.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus, 2008: *Generalized, Linear, and Mixed Models*. 2d ed., Wiley, 511 pp.

McDonald, R. E., 2011: Understanding the impact of climate change on Northern Hemisphere extra-tropical cyclones. *Climate Dynamics*, **37 (7-8)**, 1399–1425, doi:10.1007/s00382-010-0916-x.

Meehl, G. A. and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis*, S. Solomon and et al., Eds., Cambridge University Press, 747–845.

Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, **88 (9)**, 1383–1394, doi:10.1175/BAMS-88-9-1383.

Min, S.-K. and A. Hense, 2006: A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophysical Research Letters*, **33 (8)**, L08 708, doi:10.1029/2006GL025779.

Mizuta, R., 2012: Intensification of extratropical cyclones associated with the polar jet change in the CMIP5 global warming projections. *Geophysical Research Letters*, **39**, L19 707, doi:10.1029/2012GL053032.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, **117**, D08 101, doi:10.1029/2011JD017187.

Moss, R. H., et al., 2010: The next generation of scenarios for climate change research and assessment. *Nature*, **463 (7282)**, 747–756, doi:10.1038/nature08823.

Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton, and M. J. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society A*, **365**, 1993–2028, doi:10.1098/rsta.2007.2077.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:10.1038/nature02771.

Murray, R. J. and I. Simmonds, 1991: A numerical scheme for tracking cyclone centres from digital data. Part 1: development and operation of the scheme. *Australian Meteorological Magazine*, **39 (3)**, 155–166, URL `http://www.bom.gov.au/amm/papers.php?year=1991`.

Neu, U., et al., 2013: IMILAST: A community effort to intercompare extratropical cyclone detection and tracking algorithms. *Bulletin of the American Meteorological Society*, **94 (4)**, 529–547, doi:10.1175/BAMS-D-11-00154.1.

Nychka, D. and C. Tebaldi, 2003: Comments on calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the Reliability Ensemble Averaging (REA) method. *Journal of Climate*, **16**, 883–884, doi:10.1175/1520-0442(2003)016⟨0883:COCOAU⟩2.0.CO;2.

O'Gorman, P. A., 2012: Sensitivity of tropical precipitation extremes to climate change. *Nature Geoscience*, **5**, 697–700, doi:10.1038/ngeo1568.

O'Hagan, A., 2006: Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, **91**, 1290 – 1300, doi:10.1016/j.ress.2005.11.025.

Olejnik, S. and J. Algina, 2003: Generalized Eta and Omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, **8 (4)**, 434–447, doi:10.1037/1082-989X.8.4.434.

Onogi, K., et al., 2007: The JRA-25 reanalysis. *Journal of the Meteorological Society of Japan. Ser. II*, **85 (3)**, 369–432, doi:10.2151/jmsj.85.369.

Oreskes, N., K. Shrader-Frechette, and K. Belitz, 1994: Verification, validation, and confirmation of numerical models in the Earth sciences. *Science*, **263 (5147)**, 641–646, doi:10.1126/science.263.5147.641.

Palmer, T. N., et al., 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, **85 (6)**, 853–872, doi:10.1175/BAMS-85-6-853.

Parker, W. S., 2006: Understanding pluralism in climate modeling. *Foundations of Science*, **11 (4)**, 349–368, doi:10.1007/s10699-005-3196-x.

Peña, M. and H. van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *Journal of Climate*, **21 (24)**, 6521–6538, doi:10.1175/2008JCLI2226.1.

Pennell, C. and T. Reichler, 2011: On the effective number of climate models. *Journal of Climate*, **24 (9)**, 2358–2367, doi:10.1175/2010JCLI3814.1.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8441–8446, doi:10.1073/pnas.0900094106.

Pinto, J. G., U. Ulbrich, G. C. Leckebusch, T. Spangehl, M. Reyers, and S. Zacharias, 2007: Changes in storm track and cyclone activity in three SRES ensemble experiments with the ECHAM5/MPI-OM1 GCM. *Climate Dynamics*, **29**, 195–210, doi:10.1007/s00382-007-0230-4.

Räisänen, J., 2007: How reliable are climate models? *Tellus A*, **59 (1)**, 2–29, doi:10.1111/j.1600-0870.2006.00211.x.

Räisänen, J. and T. N. Palmer, 2001: A probability and decision-model analysis of a multimodel ensemble of climate change simulations. *Journal of Climate*, **14 (15)**, 3212–3226, doi:10.1175/1520-0442(2001)014⟨3212:APADMA⟩2.0.CO;2.

Räisänen, J., L. Ruokolainen, and J. Ylhäisi, 2010: Weighting of model results for improving best estimates of climate change. *Climate Dynamics*, **35**, 407–422, doi:10.1007/s00382-009-0659-8.

Randall, D. A., et al., 2007: Cilmate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, chap. 8, 589–662.

Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, **89**, 303–311, doi:10.1175/BAMS-89-3-303.

Reifen, C. and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophysical Research Letters*, **36**, L13 704, doi:10.1029/2009GL038082.

Rienecker, M. M., et al., 2011: MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of Climate*, **24 (14)**, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.

Rougier, J., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81 (3-4)**, 247–264, doi:10.1007/s10584-006-9156-9.

Rougier, J. and M. Goldstein, 2014: Climate simulators and climate projections. *Annual Review of Statistics and Its Application*, **1 (1)**, 103–123, doi:10.1146/annurev-statistics-022513-115652, `http://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-022513-115652`.

Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multimodel ensembles. *Journal of the American Statistical Association*, **108**, 852–863, doi:10.1080/01621459.2013.802963.

Rougier, J., D. M. Sexton, J. M. Murphy, and D. Stainforth, 2009: Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *Journal of Climate*, **22**, 3540–3557, doi:10.1175/2008JCLI2533.1.

Saha, S., et al., 2010: The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society*, **91 (8)**, 1015–1057, doi:10.1175/2010BAMS3001.1.

Sain, S. R., D. Nychka, and L. Mearns, 2011: Functional ANOVA and regional climate experiments: A statistical analysis of dynamic downscaling. *Environmetrics*, **22 (6)**, 700–711, doi:10.1002/env.1068.

Sanderson, B. M. and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophysical Research Letters*, **39**, L16 708, doi:10.1029/2012GL052665.

Sanderson, B. M., et al., 2008: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *Journal of Climate*, **21**, 2384–2400, doi:10.1175/2008JCLI1869.1.

Sandgathe, S., B. Brown, B. Etherton, and E. Tollerud, 2013: Designing multimodel ensembles requires meaningful methodologies. *Bulletin of the American Meteorological Society*, **94**, 183–185, doi:10.1175/BAMS-D-12-00234.1.

Sansom, P. G., D. B. Stephenson, C. A. T. Ferro, G. Zappa, and L. Shaffrey, 2013: Simple uncertainty frameworks for selecting weighting schemes and interpreting multi-model ensemble climate change experiments. *Journal of Climate*, **26**, 4017–4037, doi:10.1175/JCLI-D-12-00462.1.

Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti, 2011: Analyzing precipitation projections: A comparison of different approaches to climate model evaluation. *Journal of Geophysical Research: Atmospheres*, **116**, D10 118, doi:10.1029/2010JD014963.

Sexton, D. M. H., J. M. Murphy, M. Collins, and M. J. Webb, 2012: Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Climate Dynamics*, **38**, 2513–2542, doi:10.1007/s00382-011-1208-9.

Shiogama, H., S. Emori, N. Hanasaki, M. Abe, Y. Masutomi, K. Takahashi, and T. Nozawa, 2011: Observational constraints indicate risk of drying in the Amazon basin. *Nature Communications*, **2**, doi:10.1038/ncomms1252.

Slingo, J., et al., 2014: The recent storms and floods in the UK. Tech. rep., Met Office. URL `http://www.metoffice.gov.uk/media/pdf/1/2/Recent_Storms_Briefing_Final_SLR_20140211.pdf`.

Smith, L. A., 2002: What might we learn from climate forecasts? *Proceedings of the National Academy of Sciences of the United States of America*, **99 (suppl 1)**, 2487–2492, doi:10.1073/pnas.012580599.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, **104**, 97–116, doi:10.1198/jasa.2009.0007.

Soden, B. J. and I. M. Held, 2006: An assessment of climate feedbacks in coupled ocean-atmosphere models. *Journal of Climate*, **19 (14)**, 3354–3360, doi:10.1175/JCLI3799.1.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, (Eds.) , 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

Son, S.-W., et al., 2010: Impact of stratospheric ozone on Southern Hemisphere circulation change: A multimodel assessment. *Journal of Geophysical Research: Atmospheres*, **115**, D00M07, doi:10.1029/2010JD014271.

Spiegelhalter, D., 2006: Two brief topics on modelling with WinBUGS. *IceBUGS 2006*, Hanko, Finland, URL `http://www.math.helsinki.fi/openbugs/IceBUGS/Presentations/SpiegelhalterIceBUGS.pdf`.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde, 2002: Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639, doi:10.1111/1467-9868.00353.

Stainforth, D. A., M. R. Allen, E. R. Tredger, and L. A. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A*, **365**, 2145–2161, doi:10.1098/rsta.2007.2074.

Stainforth, D. A., et al., 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406, doi:10.1038/nature03301.

Steiger, J. H., 2004: Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, **9 (2)**, 164–182, doi:10.1037/1082-989X.9.2.164.

Stephens, M. A., 1974: EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69 (374)**, 730–737, doi:10.2307/2286009.

Stephenson, D. B., M. Collins, J. C. Rougier, and R. E. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, **23**, 364–372, doi:10.1002/env.2153.

Stocker, T. F., et al., (Eds.) , 2013: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

Stott, P. A. and C. E. Forest, 2007: Ensemble climate predictions using climate models and observational constraints. *Philosophical Transactions of the Royal Society A*, **365**, 2029–2052, doi:10.1098/rsta.2007.2075.

Stott, P. A. and J. A. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, **416**, 723–726, doi:10.1038/416723a.

Stott, P. A., J. A. Kettleborough, and M. R. Allen, 2006a: Uncertainty in continental-scale temperature predictions. *Geophysical Research Letters*, **33**, L02 708, doi:10.1029/2005GL024423.

Stott, P. A., J. F. B. Mitchell, M. R. Allen, T. L. Delworth, J. M. Gregory, G. A. Meehl, and B. D. Santer, 2006b: Observational constraints on past attributable

warming and predictions of future global warming. *Journal of Climate*, **19**, 3055–3069, doi:10.1175/JCLI3802.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, **93 (4)**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C., J. M. Arblaster, and R. Knutti, 2011: Mapping model agreement on future climate projections. *Geophysical Research Letters*, **38**, L23 701, doi:10.1029/2011GL049863.

Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A*, **365**, 2053–2075, doi:10.1098/rsta.2007.2076.

Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith, 2004: Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters*, **31**, L24 213, doi:10.1029/2004GL021276.

Tebaldi, C. and B. Sansó, 2009: Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172**, 83–106, doi:10.1111/j.1467-985X.2008.00545.x.

Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, **18 (10)**, 1524–1540, doi:10.1175/JCLI3363.1.

Thorne, P. W., et al., 2011: Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bulletin of the American Meteorological Society*, **92**, ES40–ES47, doi:10.1175/2011BAMS3124.1.

Ulbrich, U., et al., 2013: Are greenhouse gas signals of Northern Hemisphere winter extra-tropical cyclone activity dependent on the identification and tracking algorithm? *Meteorologische Zeitschrift*, **22**, 61–68, doi:10.1127/0941-2948/2013/0420.

Ventura, V., C. J. Paciorek, and J. S. Risbey, 2004: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *Journal of Climate*, **17 (22)**, 4343–4356, doi:10.1175/3199.1.

Vose, R. S., et al., 2012: NOAA's Merged Land-Ocean Surface Temperature Analysis. *Bulletin of the American Meteorological Society*, **93**, 1677–1685, doi:10.1175/BAMS-D-11-00241.1.

Watterson, I. G. and P. H. Whetton, 2011: Distributions of decadal means of temperature and precipitation change under global warming. *Journal of Geophysical Research: Atmospheres*, **116**, D07 101, doi:10.1029/2010JD014502.

Waugh, D. W. and V. Eyring, 2008: Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, **8**, 5699–5713, doi:10.5194/acp-8-5699-2008.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *Journal of Climate*, **23 (15)**, 4175–4191, doi:10.1175/2010JCLI3594.1.

Wenzel, S., P. M. Cox, V. Eyring, and P. Friedlingstein, 2013: Emergent constraints on climate carbon cycle feedbacks in the CMIP5 Earth system models. *Journal of Geophysical Research: Biogeosciences*, **Submitted**.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady, 2007: Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophysical Research Letters*, **34**, L14 701, doi:10.1029/2007GL030025.

Williamson, D., M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki, 2013: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, **41**, 1703–1729, doi:10.1007/s00382-013-1896-4.

Woollings, T., J. M. Gregory, J. G. Pinto, M. Reyers, and D. J. Brayshaw, 2012: Response of the North Atlantic storm track to climate change shaped by ocean-atmosphere coupling. *Nature Geoscience*, **5 (5)**, 313–317, doi:10.1038/NGEO1438.

Yip, S., C. A. T. Ferro, D. B. Stephenson, and E. Hawkins, 2011: A simple, coherent framework for partitioning uncertainty in climate predictions. *Journal of Climate*, **24 (17)**, 4634–4643, doi:10.1175/2011JCLI4085.1.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *Journal of Climate*, **16 (22)**, 3834–3840, doi:10.1175/1520-0442(2003)016⟨3834:IOTMST⟩2.0.CO;2.

Zappa, G., L. C. Shaffrey, and K. I. Hodges, 2013a: The ability of CMIP5 models to simulate North Atlantic extratropical cyclones. *Journal of Climate*, **26**, 5379–5396, doi:10.1175/JCLI-D-12-00501.1.

Zappa, G., L. C. Shaffrey, K. I. Hodges, P. G. Sansom, and D. B. Stephenson, 2013b: A multi-model assessment of future projections of North Atlantic and European

extratropical cyclones in the CMIP5 climate models. *Journal of Climate*, **26**, 5846–5862, doi:10.1175/JCLI-D-12-00573.1.